joy of knowledge

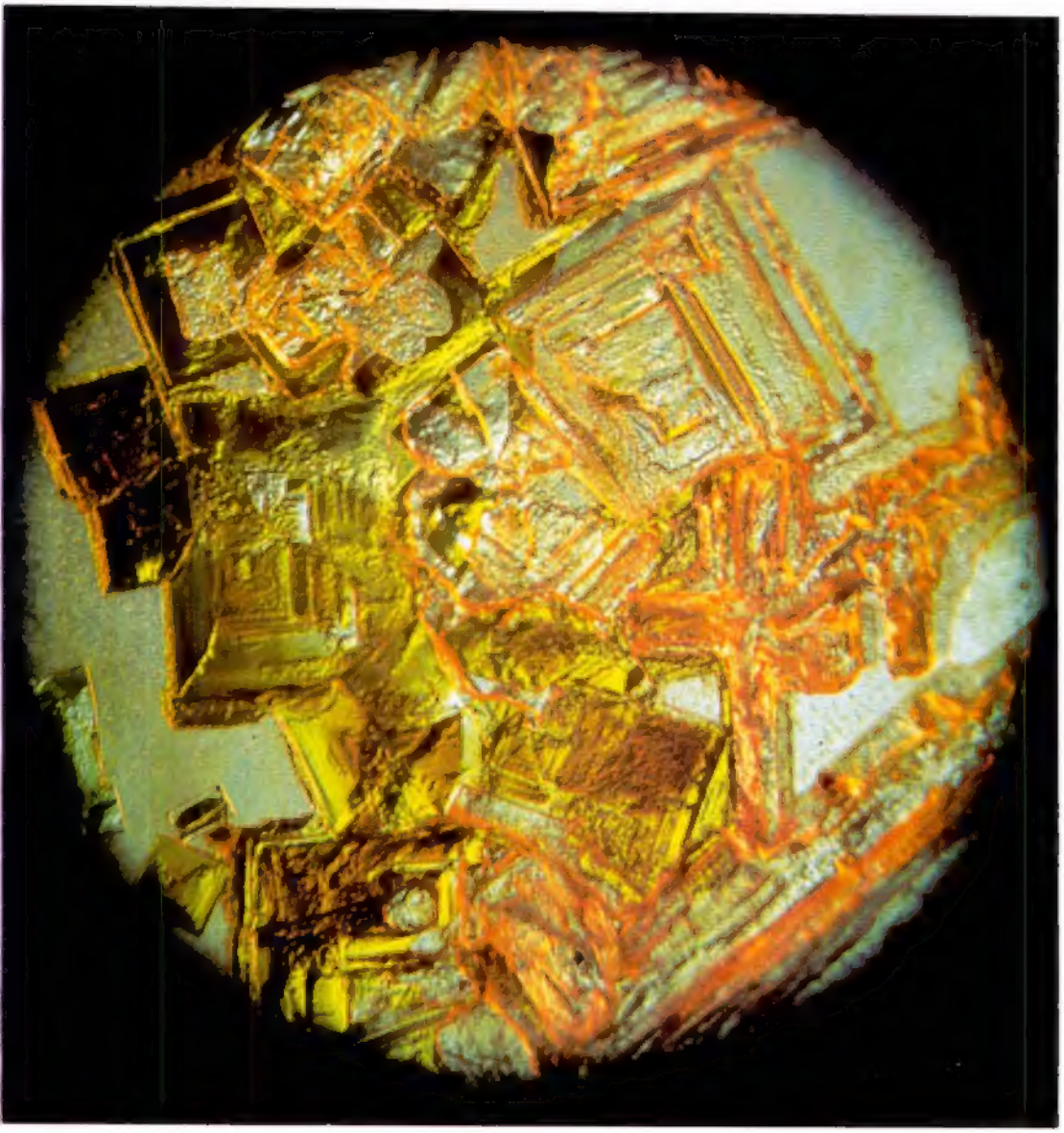
Physical Sciences

Space Sciences

Medicine & Psychology

769 – 1152

4 Physical Sciences



Prehistoric and ancient science

Nearly two million years ago the ancestors of man used stones as weapons and tools. The need to master the environment and develop tools for the purpose involved a primitive knowledge of science [Key]. Ever since that time the development of science, technology and civilization have been interrelated.

Science of prehistoric man

After perhaps half a million years, the descendants of these first tool-users had become more selective, concentrating on flint, with its useful cutting edges [1, 2]. They discovered how to make fires.

Cave paintings of 15,000 years ago indicate a primitive knowledge of animal anatomy. Some, which show prehistoric elephants with the position of the animal's heart indicated by arrows, may be a record of hunting prowess, or a form of sympathetic magic attempting to influence the hunt.

About 10,000 years ago, men began to adopt a more settled way of life. They invented a repetitive system of food production that involved the domestication of animals and the cultivation of plants. Life in such settled communities stimulated invention in various building materials, used to provide better shelter and protection. Fires were kept burning continuously for warmth and cooking and frightening off marauders, and it is possible that the women who tended the fires were the first to notice that when clay is left near a fire it hardens into pottery. When certain stones and earths were roasted in a fire, heavy liquids sometimes ran out which, when cool, hardened into useful metals.

Egypt and Mesopotamia

The people who settled in the Nile valley found themselves in an area that was exceptionally safe and fertile. They noticed that the silt brought down by the yearly floods renewed the fertility of the soil so they dug channels and built embankments to divert the fertile silt-laden flood waters onto their fields. These flood-control operations marked the invention of large-scale engineering. The Nile valley dwellers later applied engineering techniques in the construction of the pyramids [5].

It is probable that the science of geometry arose from the need to fix positions where landmarks had been washed out by annual flooding. Arithmetic was developed to calculate quantities of crops so that they could be shared among the people. Egyptian arithmetic depended on a method of doubling, much as in the operations of a modern computer. Geometrical surveying led the Egyptians to a good method of calculating the area of a circle; they assumed the circumference of a circle to be the square of eight-ninths times its diameter. They also needed to determine the seasons and the times of the Nile floods and so devised a calendar of 365 days.

The people of Mesopotamia, in the double valley of the rivers Tigris and the Euphrates, were developing in a similar way but under rather different conditions. There was little stone in their land, so they recorded information by making marks on soft clay tablets, which were then baked. The Mesopotamians introduced the idea that the value of a digit depends on its position in a number and they even solved algebraic equations. The Egyptians were primarily concerned with simple, practical calculations

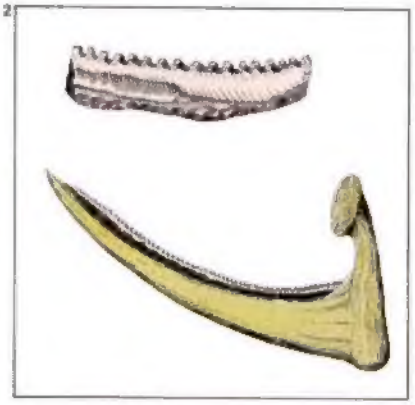
CONNECTIONS

See also
Mathematics and civilization
Asian and medieval science

1

1 Tools of the middle Palaeolithic or Mousterian stage in France date from 70,000 to 32,000 years ago. There were various human groups living in Europe in this period. They left many traces of their culture in shelters and entrances to caves. Tools are found dating from the beginning of the last period of glaciation and the scrapers and knives shown here illustrate fine stone tool technology.

2



2 Bones, antlers and wood were used by early men as raw materials for tools and weapons. A wooden tool could be given an efficient cutting edge by adding a row of suitably shaped flint slivers, as in this early Egyptian wooden sickle from about 3000–2500 BC. Pointed wooden implements could also be hardened by charring the point in a fire. Antlers were shaped by carving or heating.

3
A

B

3 The Neolithic people who occupied the region of Stoney Littleton, England in about 3000 BC, constructed large barrows for the burial of their dead [A]. A barrow is a long structure with an entrance passage crossed by transept chambers; this barrow has three pairs. The passage [B] contains a vault with corbels or projections from the walls to carry "capstones", in a manner similar to those in vaults found on islands lying off the coast of Scotland.

whereas the Babylonians were much more sophisticated, especially in connection with the science of astronomy.

One of the first Greeks to visit and study other civilizations was Thales of Miletus, who was born in about 630 BC. He returned from Egypt well versed in the techniques of Egyptian geometry. From experience in building, the Egyptians had learned that if a triangle has two sides of equal length, then the angles at its base are also equal. Thales looked for a way of proving this fact. He made two identical triangles, each with two equal sides, and found that when one was picked up and turned over, it could be laid on top of the other and fitted exactly. In this way, a mathematical proof was developed.

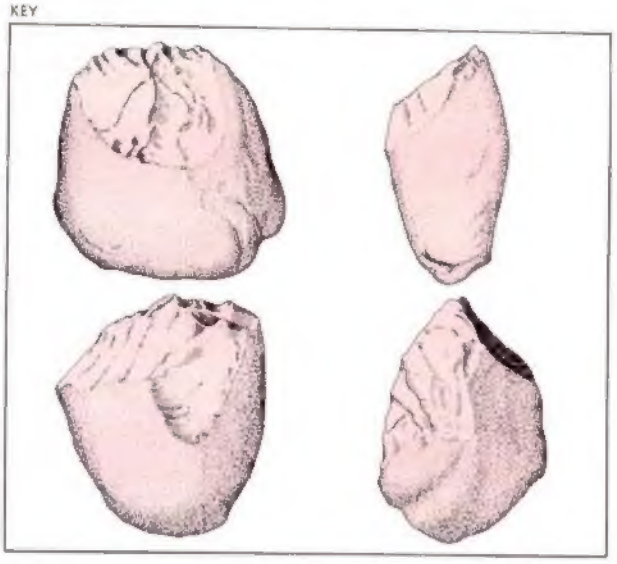
Famous Greek scientists

To Pythagoras, born about 60 years after Thales, is attributed a proof of the famous theorem that the square on the longest side of a right-angled triangle is equal to the sum of the squares on the other two sides. Pythagoras sought to explain the properties of matter in terms of numbers.

Another Greek, Euclid (born 330 BC), provided the basic principles for the teaching of classical geometry that have been used ever since. Less than 50 years later, in 287 BC, Archimedes was born in Sicily. He applied the new mathematics with extraordinary power and logic and made many inventions. He established the principle that, when a body is weighed in a liquid, its apparent loss in weight is equal to the weight of the liquid displaced; he is credited with inventing a screw for raising water [7] from one level to another; and he succeeded in launching a large ship using levers.

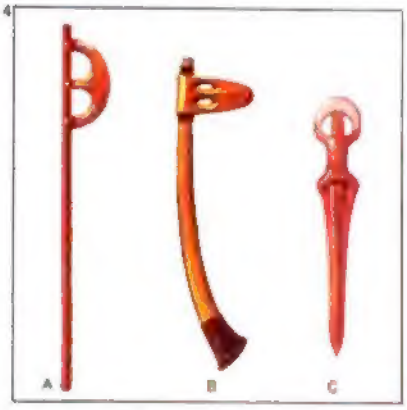
Astronomy was first placed on an adequate scientific basis by Eudoxus, who was born in about 408 BC. He showed that the motions of the Sun and planets could be explained by assuming that they moved with uniform motion in perfect circles, the centres of which are near, but not exactly at, the centre of the Earth. Later Greek astronomers arranged far more complex systems of circular motions, equalling in accuracy the work of Nicolas Copernicus (1473–1543), nearly 2,000 years later.

KEY



These Oldowan tools were first discovered by Louis Leakey (1903–72) in 1931 in the Olduvai Gorge in northern Tanzania, East Africa. They range from simple broken pebbles to chopping tools, and are 1.2 million to 1.8 million years old.

4



4 The introduction of metals gave early toolmakers much more manageable raw materials – first soft metals such as gold and copper, later bronze and finally iron. Shown here are an Egyptian eye axe [A] from Megiddo (*c.* 1900 BC), an Egyptian duck-bill axe [B] from Ugarit (*c.* 1800 BC) and an Egyptian bronze dagger [C] of the Hyksos period (*c.* 1650 BC). But iron made the best cutting edge.

5

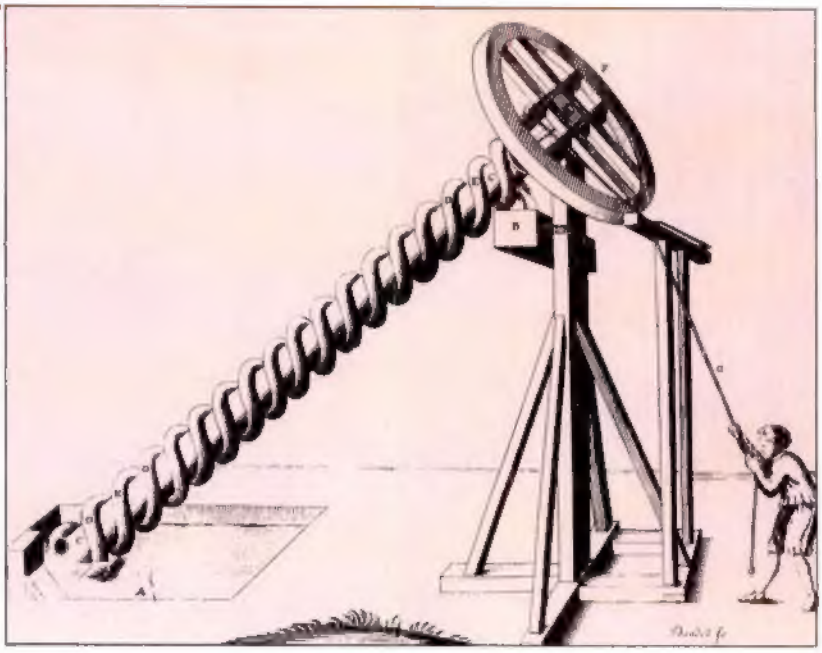
5 The Great Pyramid in Egypt was erected by command of the pharaoh Khufu (in Greek, Cheops) of the 4th Dynasty, about 3000 BC. It contains 6.5 million tonnes of limestone. A primary purpose of the pyramids was to provide grand tombs. They may also have had other purposes. The whole group of the major Egyptian pyramids was built within little more than a century. They involved an enormous concentration of labour and it has been suggested that their construction may have provided a convenient means of organizing the whole population of Egypt, creating a centralized state – hence the short building period.

6

6 The Mesopotamians, lacking the Egyptians' papyrus for writing, made records on clay tablets. Characters were formed by pressing the wedge-shaped end of a stylus into the soft clay and a permanent record could be obtained by baking the clay in a fire until it became hard. From its wedge shape, the writing is known as cuneiform. The associated number system used strokes that, to our eyes, faintly resemble the Roman numerals of 30 centuries later. A numeral's position determined its value.

7 An Archimedean screw is used for raising water. Archimedes (287–212 BC) was a master both of the most refined mathematics and of practical invention. It is said that he invented the screw for raising water to assist irrigation in Egypt. It consists, in principle, of a wedge that can exert sustained pressure in a particular direction by being continuously revolved. In the machine shown in this illustration the water is raised from the sink on the left. When the curved pipe is rotated by pulling the rope, the water is pushed up inside it and delivered into the tank at the top on the right. Screws were not known before Archimedes: those for fastening objects probably arose from his device.

7



Asian and medieval science

By the fifth century AD the long-ailing Roman Empire lost control of Western Europe. It was overcome by its excessive size, by the "softness" of its citizens, and by a population explosion in Asia that propelled vigorous new peoples against Rome's extended frontiers. At the same time the Eastern Roman Empire, whose capital was at Byzantium (later Constantinople and now Istanbul), flourished until a new attack from the east in the seventh century.

International influences

The Arabs emerged from Arabia as the followers of the new prophet Mohammed (AD *c.* 570–632), a trader from Mecca. Within about 100 years they captured much of the Middle East and North Africa and invaded Spain and even France [2]. These new conquerors had no culture of their own, but they borrowed learning from Syriac, Greek, Indian and other peoples whom they encountered through their conquests and travels. They became the founders of the internationalism that is one of the most striking features of science. They aimed at all-embracing knowledge and perhaps Avicenna (980–1037) came nearest to attaining it.

The Muslims had a strong trading tradition. Like other peoples, they were interested in the exact calculation of shares in goods and the allocation of family inheritances. They assessed the declining value of female slaves in much the same way as cars are depreciated today. When they invaded India, they discovered Indian mathematics.

The Indians introduced the number system now universally in use, together with the zero symbol and decimals. Their work became known to Al-Khwarizmi (780–*c.* 850), the greatest of the Arab mathematicians. He was librarian to the caliph Al-Mamun (786–833) in Baghdad and published in 830 his treatise on *Al-jabr wa'l muqabala*, from which the word algebra is derived. He studied various classes of quadratic (second order) equations and called the unknown quantity to be calculated "the root". Knowledge of mathematics in medieval Europe was based mainly on Latin translations of his works. One of his most famous successors was the Persian poet Omar Khayyám (*c.* 1048–1122), who dealt with special classes of cubic equations.

The Arabs devoted much attention to pharmacy [5] and to astronomy. They calculated elaborate trigonometrical tables, which were used to determine the exact times of prayers and to navigate the Indian Ocean. Córdoba in Spain became the most advanced intellectual centre in Europe. The dependence of medieval Europe on Arab knowledge is illustrated by the example of Adelard of Bath, who went to Córdoba disguised as a Muslim student in 1120. He returned to England with a copy of Euclid's book, which served as a mathematical text for feudal Europe for the next four centuries.

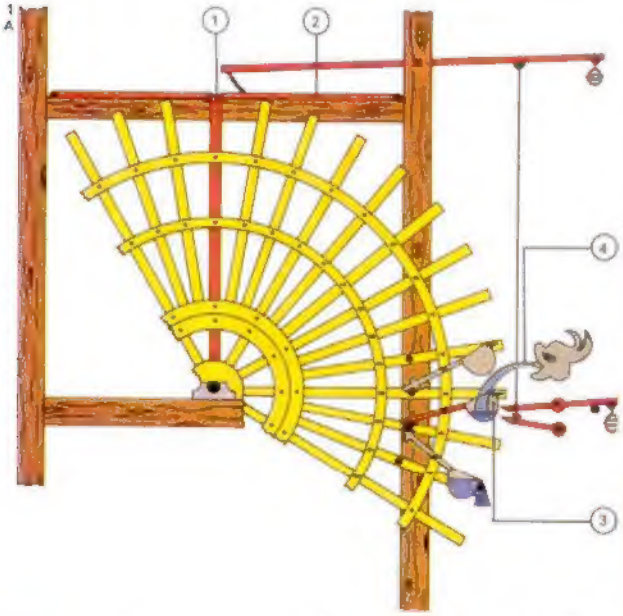
Chinese science and technology

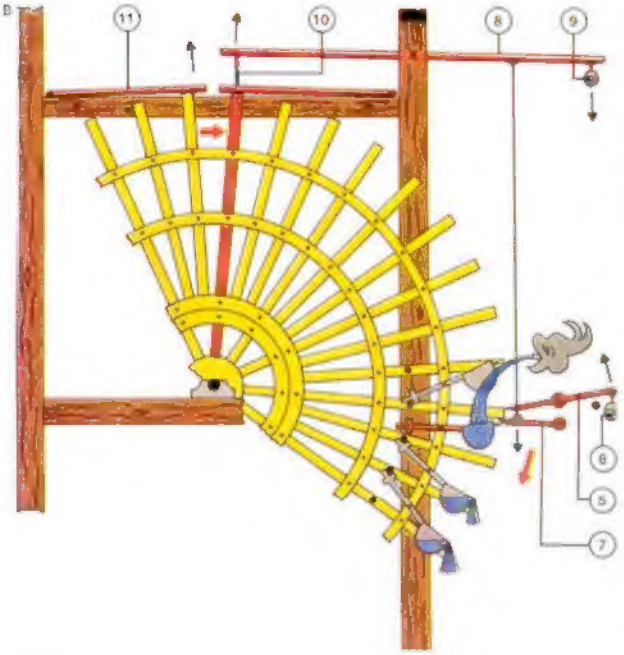
The Arabs also brought knowledge of Chinese inventions and discoveries to Europe, including gunpowder, the magnetic compass, printing with movable type and an efficient horse harness.

The Chinese invented an escapement mechanism for a water clock [1], generally credited to Yi Hoing, in 725. It enabled them

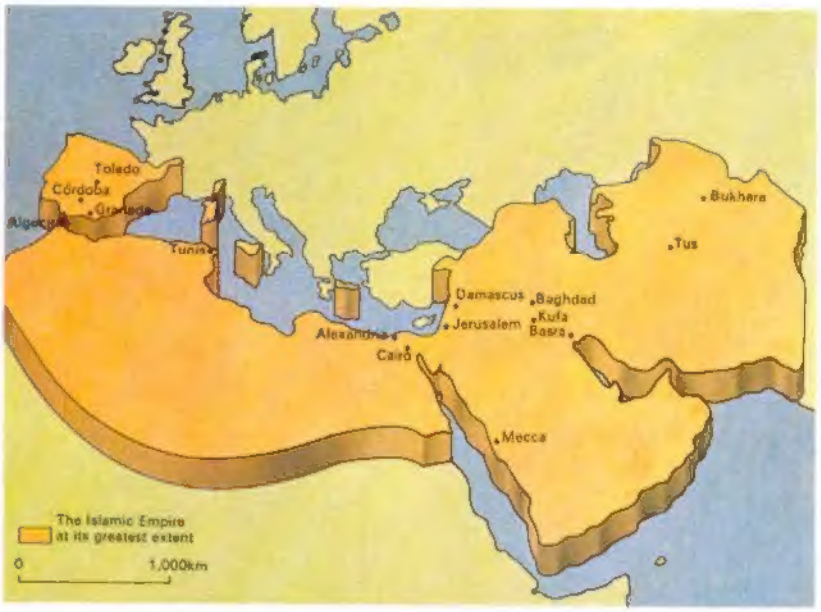
CONNECTIONS

See also
Mathematics and civilization
Finding unknown quantities: algebra
Magnets and magnetism

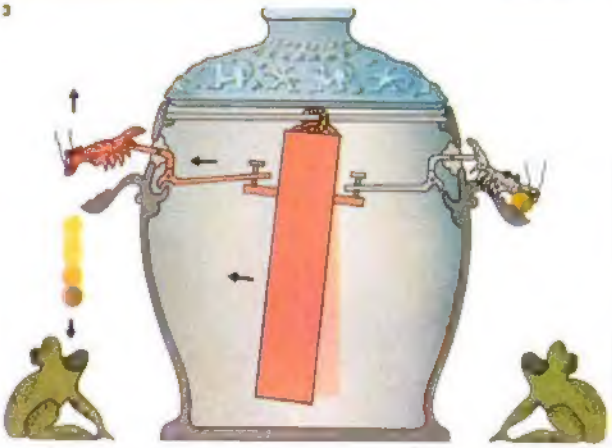




1 This clock escapement mechanism was invented in China in the 8th century. [A] A spoke [1] is arrested by a lock [2] while a scoop [3] fills with water [4] from a tank at a constant rate. The lock is released [B] when the filling scoop trips a checking fork [5], overcomes its counterweight [6] and trips a coupling tongue [7], which pulls down an upper lever [8] with its own counterweight [9]. This jerks a chain [10], freeing the lock and allowing the wheel to swing clockwise until the lock drops again, arresting the following spoke which is also held steady by a ratchet lock [11].



2 The Islamic Empire, by the 8th century, stretched from India to the Pyrenees, making major contributions to mathematics and chemistry and laying the basis for the international spread of learning.



3 A seismograph built in 132 by the Chinese scientist Chang Heng was a vase with a ring of holes round the rim. Metal balls lightly held in each hole fell into receptacles below when there was an earth tremor. It indicated the direction of the tremor according to which balls fell and which did not. In a modern attempted reconstruction of the internal mechanism, a pendulum with arms and cranks governs the motion of the balls.

to build the first accurate mechanical clocks. Chinese science also produced the first seismograph, for detecting earthquakes [3], built by Chang Heng in 132. In 1054 they observed the great new star – a nova, the parent of the Crab Nebula – later to become one of the most important objects in the development of modern radio astronomy.

The Arabs performed a unique service as world informants on ancient and contemporary science. They introduced the ideas of India and China to the Western world. But they were not the founders of modern science, for this arose in Europe.

Medieval science

The Roman Empire in Europe disintegrated into a multitude of individual strongholds of local military chieftains. Slaves became bandits, or peasants tied to the land, but no longer slaves. So the feudal period began. People who settled around the strong-points came to be known as the bourgeoisie, because they lived outside the castle or burg. Many of them were craftsmen. They were dependent on what they could learn from the Arab encyclopaedists but they looked at the old knowledge in a more individual way. One such man who played a significant role in the advance of medieval science was Leonardo Fibonacci of Pisa (*c.* 1180–1250). His father was employed on the Barbary coast and there Leonardo learned the Arabic language and arithmetic. On his return to Pisa he introduced Arabic numerals into Europe.

The most eminent of English medieval scientists was Roger Bacon (1214–94). He proposed combinations of lenses for telescopes and microscopes and may have been the first to suggest spectacles [4].

Roger Bacon said that the only man he knew who was to be praised for his experimental science was Petrus Peregrinus of Maricourt, who published a treatise on magnetism in 1269. He explicitly pointed out the importance of manual skill in science. He was one of the forerunners of modern science, in which experiment and theory are equally balanced. This became possible only through the emancipation of the craftsman in medieval Europe, a change that was to lead to the Renaissance.

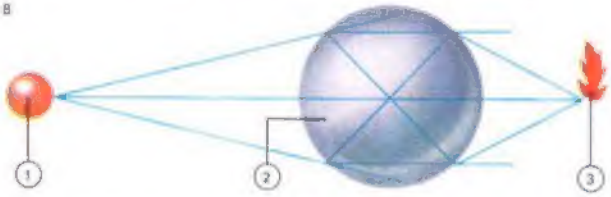
KEY

An Arab map of remarkably modern graphic style shows the seas and land masses known to medieval Arab geographers. The concept of a round world with an encircling ocean and the territorial extent of the map indicates the importance of the Arab contribution to man's knowledge.

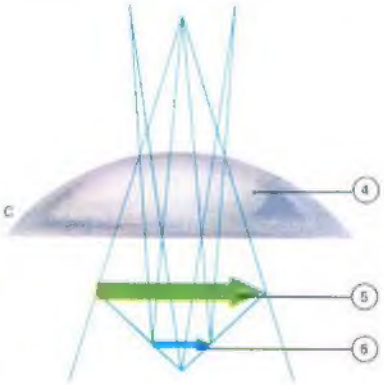
4 A

4 Roger Bacon [A] was not personally an experimenter or mathematician but he realized the importance of experiment and mathematics for the advancement of science. His imagination enabled him to make a remarkable collection of scientific suggestions, culled from many sources and ranging from vague hints to clear diagrams. He gave substantially correct optical explanations of [B] why a spherical flask of water [2] acts as a burning glass [3] to concentrate the rays of the sun [1] and [C] how a convex lens [4] produces a magnified image [5] of an object [6] beneath it.

B



C

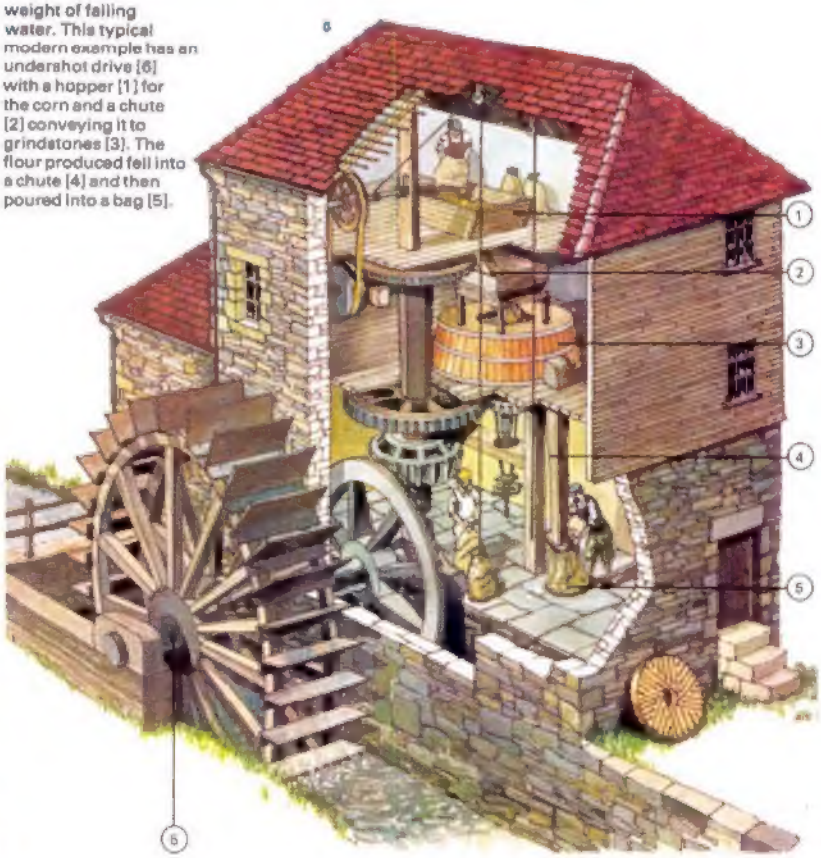


5 A typical pharmacy shop of the Middle Ages was based on Arab traditions. The Arabs brought knowledge of Persian and Indian drugs and spices such as camphor, cloves, cassia, nutmeg and senna. Their influence on European pharmacy was exerted chiefly through Benedictine monks. In the 14th century pharmacy, medicine, chemistry and the grocery business were combined and apothecaries and grocers organized themselves in guilds.

5

6 Corn mills began with the Romans who spread techniques that helped them to exploit their empire. They built undershot and overshot water mills, one driven by the momentum of flowing water and the other by the weight of falling water. This typical modern example has an undershot drive [6] with a hopper [1] for the corn and a chute [2] conveying it to grindstones [3]. The flour produced fell into a chute [4] and then poured into a bag [5].

6



Alchemy and the age of reason

Men gained some practical knowledge of the working of materials in early times but work in the crafts (what is now called technology) was regarded as a lowly pursuit. One reason for this attitude was the disagreeable working conditions associated with them. An Egyptian scribe of 1500 BC noted that the metalworker "stinks like fish-spawn".

Alchemy and the Renaissance

In the second century AD, Diocletian (245–313) ordered that all books on the working of gold, silver and copper should be destroyed to prevent counterfeiting and inflation. The effect was to reduce rational research on practical problems and to increase interest in magic as a method of transmuting base metals into gold. The centre of the development was Alexandria and the Arabs called the new science *alchemy* after Khem, or "black", the name given to Egypt because of its black earth.

The Alexandrians invented apparatus for heating, melting, filtering and distilling substances. They introduced the glass flasks and retorts still typical of chemical laboratories. The Arabs [1] adopted, extended and transmitted these advances. Their greatest chemist was Jabir ibn Haijan, or Gebir (*c.* AD 721–817), who worked on the transmutation of metals and propounded a theory of their constitution that was not completely superseded until the eighteenth century. Besides being familiar with chemical operations such as crystallization, solution and reduction, he attempted to explain them. His most useful discovery was nitric acid.

Modern science was founded during the Renaissance in the urban society of Italian cities where craftsmen became emancipated and even famous. The supreme example was Leonardo da Vinci (1452–1519) who knew little Latin and no Greek but analysed technical processes scientifically.

Copernicus and Galileo

Nicolas Copernicus (1473–1543) was a Polish–German scholar who studied at Cracow and Bologna in the 1490s. He noted from astronomical references in Latin and Greek literature that Heraclides (388–315 BC) had assigned a motion to the earth "after the manner of a wheel being carried on its own axis". Copernicus found "by much and long observation" that a consistent account of the movements of the planets could be given on the basis that the earth revolves around the sun. The account of this in Copernicus' treatise *De Revolutionibus Orbium Caelestium* was published in 1543 when he was on his deathbed. The Copernican theory [Key] is perhaps the most important scientific theory in history for it changed man's conception of his place in the universe. Formerly man had believed that the universe revolved around the earth and himself; now he realized that man was but a minute incidental speck in a universe of almost inconceivable vastness.

The Renaissance effort in science was completed by Galileo Galilei [2] who cleared the way for modern science. Copernicus discovered how the Solar System works but Galileo gained the first precise knowledge of how things on the earth move. He was born in Pisa in 1564, the same year as Shakespeare, and died in 1642, the year in which Isaac Newton was born. He went to the local

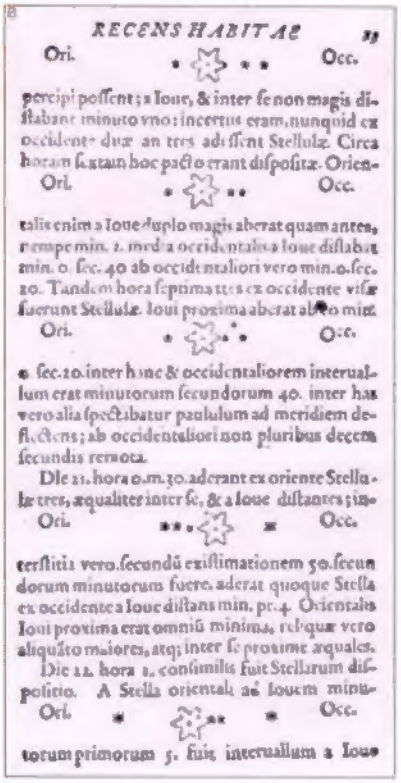
CONNECTIONS

See also
Asian and medieval science
Lines and angles
Trigonometry
Beyond the atom
Classification of elements

1

1 Alchemical processes are represented by the figures on this Arab manuscript reflecting cross-cultural influences. The Arabs gained their main introduction to alchemy through Alexandria and spread it to Western Europe in about the eleventh century. Although they improved on the experimental techniques of the Alexandrians they did not escape the influence of their mystical theorizing which was based on animistic beliefs in objects possessing souls.

B



RECENS HABITAE

Ori. Occ.

percipi poſſent; a Ioue, & inter ſe non magis diſtabant minuto vno: incertus eram, nunquid ex occidente duæ an tres adeſſent Stellulæ. Circa horam ſextam hoc pacto erant diſpoſitæ. Orien-

Ori. Occ.

talis enim a Ioue duplo magis aberat quam antea, nempe min. 2. media occidentalis a Ioue diſtabat min. o. ſec. 40 ab occidentali vero min. o. ſec. 20. Tandem hora ſeptima tres ex occidente viſæ ſuerunt Stellulæ. Ioui proxima aberat altero min.

Ori. Occ.

o. ſec. 20. inter hanc & occidentaliorem interuallum erat minutorum ſecundorum 40. inter has vero alia ſpectabatur paululum ad meridiem deflectens; ab occidentaliori non pluribus decem ſecundis remota.

Die 21. hora o. m. 30. aderant ex oriente Stellulæ tres, æqualiter inter ſe, & a Ioue diſtantes; in-

Ori. Occ.

terſtitia vero ſecundũ exiſtimationem 50. ſecundorum minutorum fuere. aderat quoque Stella ex occidente a Ioue diſtans min. pr. 4. Orientalis Ioui proxima erat omniũ minima, reliquæ vero aliquãto maiores, atq; inter ſe proxime æquales.

Die 22. hora 1. conſimilis fuit Stellarum diſpoſitio. A Stella orientali ad Iouem minu-

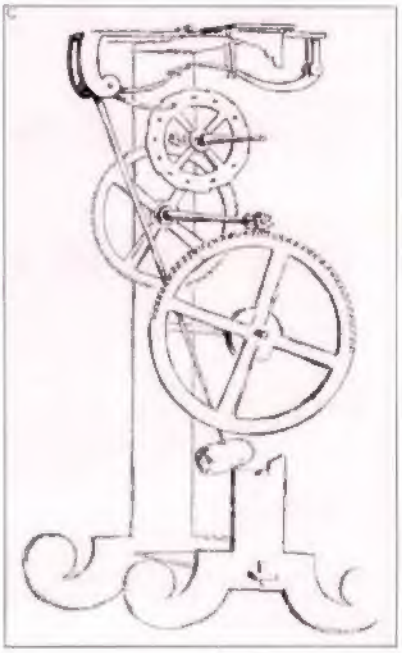
Ori. Occ.

torum primorum 5. fuit interuallum a Ioue

2 A

2 Galileo [A] discovered the constancy of the swing of a pendulum, which he later adopted into a pendulum clock [C]. His development of the telescope led to original observations on the planets, including the discovery of the satellites of Jupiter, first described in a pamphlet of 1610, a page of which [B] is shown here. The Church imprisoned him for refuting its divine knowledge.

C



3

3 The astrolabe, invented by the Greek Hipparchus around 150 BC and developed by Arab scholars and astronomers, was introduced to Europe in about the tenth century. Two metal discs bore projections of the celestial and terrestrial spheres. A rotating arm on the back enabled the user to set the inclination of an object from the horizon and to calculate various angles. A forerunner to the sextant, it was a navigational aid whose uses included finding angles of latitude and the time.

university. According to legend, a swinging lamp in the cathedral attracted his attention. He noticed that the time of the swing was independent of the size of the swing (its amplitude). When he arrived home he checked the fact with a bullet and a piece of string. He later used this fundamental property of a pendulum in designing a pendulum clock.

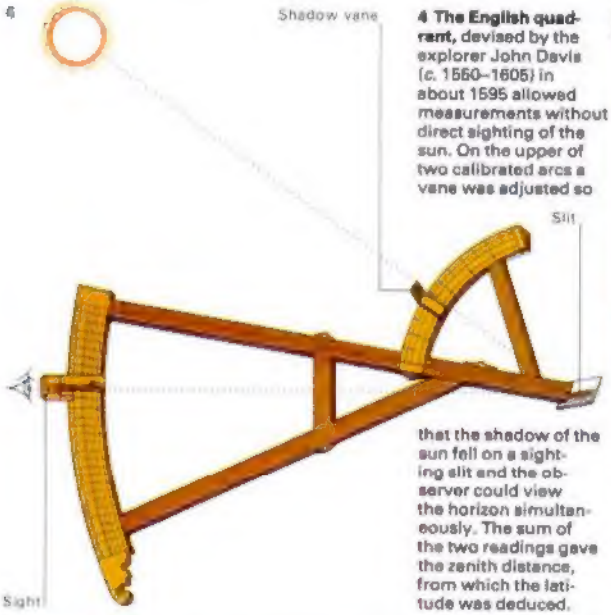
When Galileo was appointed professor he was obliged to teach Aristotelian science. This caused him to make a careful study of Aristotle's ideas, especially those on the motion of objects. Aristotle (384–322 BC) based his theory on the assumption that objects fall with a speed proportional to their weight. Galileo devised experiments to measure exactly how fast objects do fall and found that all that fall freely do so at the same speed. He made many other discoveries. The results of his application of the telescope to astronomical observation were particularly spectacular evidence that Aristotle's picture of the universe was incomplete and mistaken. Galileo did not fully grasp the theological implications of this.

Galileo's demonstration that the movement of objects could be exactly determined by a combination of experiment and mathematical reasoning was extended by Isaac Newton (1642–1727). Newton [5] showed that all the then known physical aspects of the universe and nature could be completely described by mathematical theory utilizing laws consistent with experience. Newton's account, in his *Philosophiae Naturalis Principia Mathematica* (1687), is possibly the greatest single intellectual effort yet made.

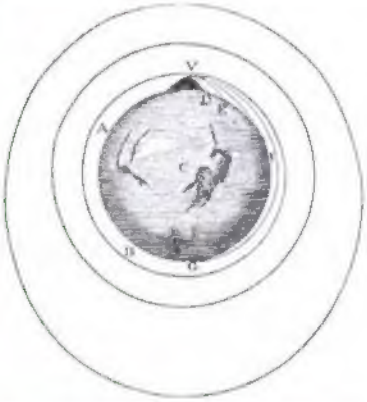
The Age of Reason

Newton's achievements increased confidence in the power of human reasoning. They had a particularly striking effect in France where Pierre Laplace (1749–1827) and Joseph Lagrange (1736–1813) extended the Newtonian theory and its supporting mathematics. The new confidence in experiment and calculation developed also in other sciences. Antoine Lavoisier (1743–94) revolutionized chemistry, dispatching the magical and mystical remnants from it and in doing so laying the foundations of modern chemistry [6].

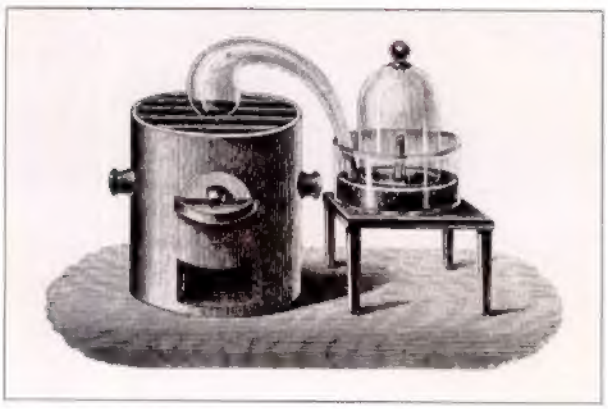
Copernicus pictured the planets as revolving around the sun in a complicated pattern of circular motions, the basis of which is shown here. His theory did not give more accurate planetary predictions than Ptolemy's of AD 140, but it was a triumph of ideas that showed man in his true place in nature. Proof that the planets revolve in ellipses came in 1609 from the German astronomer Johannes Kepler (1571–1630).



4 The English quadrant, devised by the explorer John Davis (c. 1550–1605) in about 1595 allowed measurements without direct sighting of the sun. On the upper of two calibrated arcs a vane was adjusted so that the shadow of the sun fell on a sighting slit and the observer could view the horizon simultaneously. The sum of the two readings gave the zenith distance, from which the latitude was deduced.



5 Isaac Newton [A] calculated the speed at which a body projected horizontally from the top of a mountain would leave the earth and begin revolving around it. His diagram of the path of an artificial satellite [B] was published in 1728, the year after his death. Newton's genius emerged while he was at Cambridge University. Within a few years he had laid the basis of differential and integral calculus, elucidated the nature of light and colour and had begun to explore the usefulness of mathematical analysis to physical theories. His main achievements in physical science were expressed in his *Principia* (1687) and *Opticks* (1704).



6 Antoine Lavoisier [B] founded modern chemistry by means of experiments leading to his theory about the nature of combustion. This had previously been ascribed to the transfer of a substance called phlogiston, the main agent of chemical change sometimes released as fire. Lavoisier heated mercury and air in a flask with a curved neck, which enabled him to measure exactly the decrease in volume of gas and gain in weight of mercury during 12 days' heating. By means of this apparatus (illustrated here from Lavoisier's own diagram [A]) he showed that the changes could be explained completely in terms of the active constituent of the air discovered by Joseph Priestley (1733–1804), to which Lavoisier gave the name oxygen. The idea of phlogiston was superfluous and chemistry could develop as a rational science based entirely on quantitative measurements. Experiments gave rise to theories that were then tested by other experiments.

Mathematics and civilization

Mathematics is a continuously expanding system of organized thought. It is employed in science, technology, art, music, architecture, economics, sociology, sport – in fact, in almost every aspect of human activity – and has influenced, and often determined, the direction of philosophical thought concerned with man and his universe. Throughout history mathematics has not only reflected developments in civilization but also made a major contribution to those developments.

Algebra, geometry and calculus

There are three major aspects of mathematics. The assembling and combining of sets of objects led to concepts of number [1], computation and algebra. Concern with the measurement of time and space led to geometry, astronomy and chronology. The struggle to understand ideas of continuity and limit led to mathematical analysis and the invention of calculus in the seventeenth century. These three aspects of mathematics overlap considerably. There are now, for example, algebras of sets, vector algebra and algebraic geometry, and a host of specialities that employ the concepts of other fields of study.

Everything natural or man-made has a structure comprising elements that are related in some special way [Key]. A rock crystal, a plant [6], a spaceship and a political system each has a structure, the study of which is mathematical. Mathematics is the result of the thought process known as abstraction, in which activities related to a physical structure can be organized in such a way that the physical structure is replaced by a mental one, an abstract mathematical model. The power of mathematics is further demonstrated when abstract concepts, such as those of number and space, can be represented by concrete symbols, which may be algebraic, geometric or graphical [3].

Mathematics can be described as a form of inquiry made according to defined rules for drawing conclusions from accepted mathematical truths. History shows, however, that mathematics is also a field of creative activity employing great flights of intuition and imagination [7]. The driving force for the creativity is usually the need man has to solve the problems of his society. But the motivation may also simply be the challenge of intellectual activity for its own sake.

The first mathematicians

All primitive civilizations developed concepts of number and measure as soon as trade progressed beyond the process of barter. Almost 6,000 years ago the Sumerians were using a numeration system based on 10 (denary system) as well as one based on 60 (sexagesimal system). The sexagesimal system still survives in the measure of time and rotation, reflecting the Babylonian preoccupation with the motion of the Sun, Moon and other planets and their influence on man.

The knowledge acquired became not only a religious force but also solved basic problems of agriculture and social organization. The flooding in Babylon and Egypt demanded seasonal surveys of land, the techniques of which led to geometry. Political, commercial and religious pressures to build palaces, ships, temples and tombs stimulated the further development of

CONNECTIONS

See also
The grammar of numbers
The language of numbers
Finding unknown quantities: algebra
Lines and angles: trigonometry
Shape and symmetry
Asian and medieval science
Prehistoric and ancient science

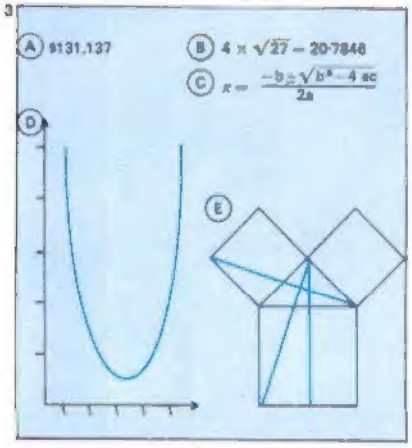
1

1 The concept of number is fundamental to mathematics. It probably developed originally out of the need for farmers to count their animals and produce. Numbers also led to money systems, making buying and selling possible.

2

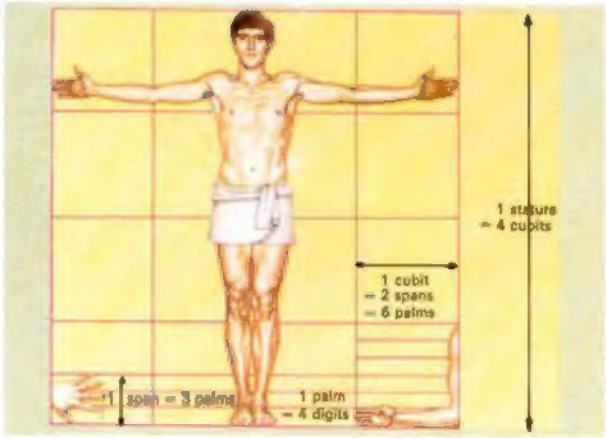
2 Stonehenge was built in the Bronze Age as a sort of calendar, which probably also had a religious significance. The positions of the stone blocks can be used to measure the movements of the Sun and Moon and to "predict" eclipses.

3



3 Mathematics has generated its own "language". Numbers are themselves shorthand forms of words and, linked with units, define exact amounts or measurements [A]. Other symbols stand for operations such as multiplication and square roots [B]. In algebra letters often stand for unknown quantities, as in this formula [C] for finding the solutions to a quadratic equation. A graph [D] can "draw" algebraic functions. Pythagoras created his own geometrical conventions [E].

4



4 Man probably first counted on his fingers and sized objects in terms of his own body. This diagram shows some of the ancient units of length. "Body units" are still used in some countries today. A hand, equal to 4in (about 10cm), is a standard unit for measuring the height of horses and in North America and Britain a foot – 12in (30.5cm) – is still used in measurement as a unit of length. The metric system is now the most widely accepted system of measurement.

geometry. At the same time, astronomy regulated social and religious events and thus served the political ends of ruling priests.

The Greeks established mathematics as a rigid study, placing mathematical argument on a logical basis so that propositions, previously not self-evident, could be deduced from basic assumptions. Euclid's *Elements*, produced in about 300 BC, was a prime example of this approach and dominated geometric thinking for 2,000 years. The Greeks saw beauty in number and shape and their excitement with the Golden Ratio [5] manifested itself in their art and architecture and has been echoed by later civilizations in such places as Notre-Dame in Paris, the architecture of Le Corbusier and the United Nations building in New York.

Every civilization has demanded systems for measuring and each new method has borrowed ideas from previous ones. As civilizations expanded, their influences and trade spread, and the need for standardized units increased. The earlier systems were all based on convenience, so that parts of the body were used for measuring length [4], the working capacity of oxen for area, stones for weight, skins for volume. Each society learned to standardize; in 1791 the French devised the metric system based on the metre, one ten-millionth of the Earth's quadrant (a quarter of the circumference), a distance calculated from an actual survey. International trade has now forced most of the Western world to adopt the metric system of measurement.

The heritage of numbers

Mathematics resembles a living organism in that its growth is affected by the environment in which it lives. The golden age of Greece produced mathematical beauty that afterwards lay dormant for centuries. The Romans used earlier mathematics but solved no new problems. Not until the sixteenth century was there another great advance. Today the whole world is experiencing change at a pace unequalled in the past. This is mirrored in the development of new mathematics and its applications in solving the problems of science, technology, industry and commerce [8] peculiar to the late twentieth century.

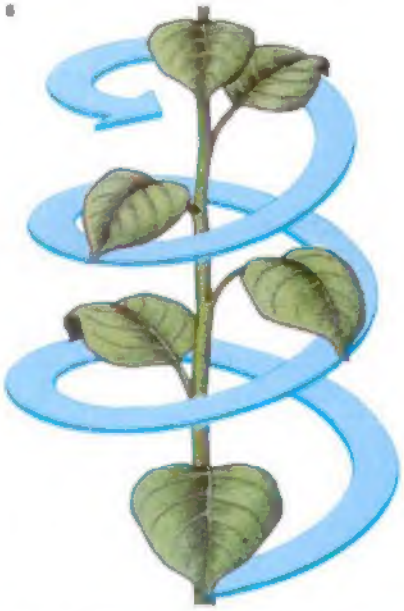
KEY

Everything on earth, from the atoms in this crystal to the leaves on a tree, consists of individual components. Mathematics seeks to establish the relationships existing between such components.

5

5 Greek mathematicians extended their logical thinking into the arts, establishing mathematical relationships in music and art. The Golden Ratio (approximately 1.618) was to the Greeks a pressing proportion, incorporated here in the Parthenon (built 447–432 BC).

6



6 Fibonacci ratios are elements in the series 1/1, 2/1, 3/2, 5/3, 8/5, 13/8 and so on. These values approach the Greek Golden Ratio. Both the numerators and the denominators in the series are formed by adding consecutive members of the series. These ratios occur in nature; a spiral following leaves on this stalk has gaps and turns in the ratio of 5/3.

7

7 The Grand Canal at Venice was a favourite subject of the Venetian painter Canaletto, whose real name was Giovanni Canal (1697–1768). Renaissance painters studied perspective and so laid the foundations of projective geometry in mathematics, map-making and the draughtsmanship used in architecture and engineering, enabling a three-dimensional object to be represented in two dimensions.

8 An electronic calculator is a modern machine for "doing sums". It and the much more complex digital computer have replaced earlier calculating devices such as the mechanical calculator, the slide rule and, oldest of all, the abacus.

8

The grammar of numbers

People use arithmetic so frequently in everyday life that they hardly ever think about it. Yet every time a woman buys something and counts her change [Key] she uses the basic concepts of addition and equality, ideas in use since trading began.

Basic rules of arithmetic

The four main types of calculations are addition, subtraction, multiplication and division. They are carried out following basic laws – most of which are merely statements of common sense. The commutative law holds for both addition and multiplication. It simply states that the sum of seven and two $(7+2)$, for example, is the same as the sum of two and seven $(2+7)$. In other words, the order in which numbers are added does not matter. The same is true of multiplication: $4\times3 = 3\times4$ or, in general terms, $a\times b = b\times a$.

The associative law is an extension of this idea and states that, in adding or multiplying a series of numbers, the order of addition or multiplication does not matter [1]. Using symbols to stand for any numbers, $(a+b)+c = a+(b+c)$, or $(a\times b)\times c = a\times(b\times c)$.

The distributive law states that if two numbers are to be added together and the sum multiplied by a third number, the same result is obtained if each of the first two is first multiplied by the third and the two products added. This law is easier to state using symbols: $(a+b)\times c = (a\times c)+(b\times c)$, and is made clear by an example: $(5+7)\times3 = (5\times3)+(7\times3) = 36$.

Multiplication is equivalent to repeated addition. For instance 7×5, is a shorthand way of writing $7+7+7+7+7$. People learn multiplication tables because it is quicker to apply them than to add columns of figures. Electronic calculators and computers, renowned for their speed and accuracy, cannot multiply; they work by successive addition, but do so extremely quickly.

Just as subtraction is the reverse of addition, so division can be regarded as the reverse of multiplication – a repeated subtraction [3]. This is the method employed in doing "long division" sums. Often it is not possible to subtract successively one number from another an exact number of times – there is generally something "left over", called the remainder. For example, 380 divided by 70, is 5, and the remainder is 30.

Squares and square roots

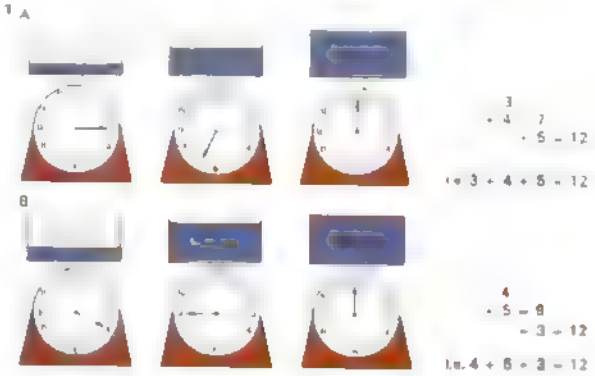
When a number is squared, it is multiplied by itself (the area of a square is the length of one side multiplied by itself). Three squared (written 3^2) equals 9. The reverse operation is called taking the square root: what number multiplied by itself makes a given number? Squaring a whole number (integer) gives an integer result, but taking the square root of a whole number often does not. And, as the Greek mathematician Pythagoras and his co-workers discovered, there is not always a rational number (expressible as the ratio of two integers) that when squared will equal a particular integer. The square root of 4 is 2 (both integers), but the square root of 2 is somewhere between 1.4142 and 1.4143. The square root of 2 cannot be computed exactly and is called an irrational number.

Fractions, proportions and ratios

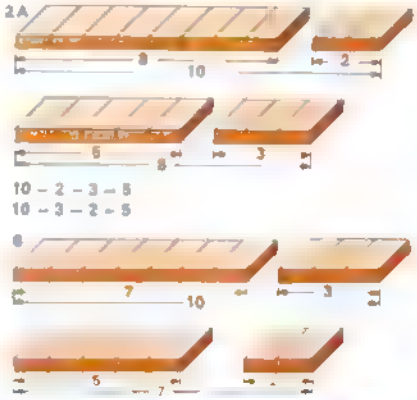
Three-sevenths is written as 3/7, meaning 3 divided by 7. It is a fraction (from the Latin

CONNECTIONS

See also

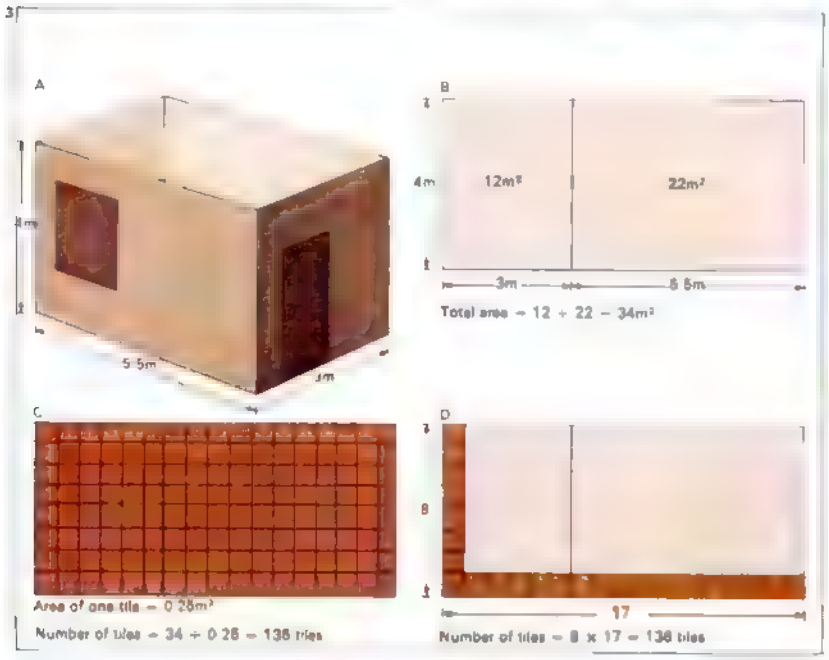


1 Addition is associative – that is, a series of additions can be carried out in any order without affecting the result. This diagram shows the effects of successively weighing (A) 3, 4 and then 5 units of a substance on a spring balance and (B) weighing 4, 5 and then 3 units. In both cases the total weight – the sum or the additions – is 12 units. As in many other mathematical laws, this is applied common sense.



2 A series of subtractions can also be carried out in any order. Starting with a piece of wood 10 units long (A), we can cut off first 2 units and then remove a further 3 units from the 8 remaining (so completing the sum 10 – 2 – 3 = 5). Or (B) we can remove first 3 units and then cut off another 2 units from the original 10. This time the subtraction sum is 10 – 3 – 2 = 5 but the result is exactly the same.

3 Multiplication and division are needed to solve many every day problems. A man wants to tile the two plain walls of a room (A), which is 5.5m long by 3m wide and 4m tall. Using tiles 0.5m square. The walls can be drawn (B) as two areas of 22m² and 12m², giving a total area of 34m². A single tile 0.5m by 0.5m has an area of 0.25m². The number of tiles required (C) can be found by dividing the area of one tile (0.25m²) into the total area to be covered (34m²), giving the result 136 tiles. The same problem can be tackled another way (D). If the whole area to be tiled is considered, it measures 8.5m by 4m. The long side will accommodate 17 half-metre tiles and the short side only 8 tiles. The total number of tiles required is therefore 17 × 8 = 136, the same result as before but without calculating areas.



4 Dividing quantities into equal parts is a method of forming them into fractions. An athlete such as a pole-vaulter intuitively judges his run-up by dividing it into an equal number of paces so that the pole is in exactly the correct place for the jump – he cannot make half a pace. The same is true when we speak of a bottle being half full or say that we have read a third of a book. In a fraction such as three-quarters, written 3/4, 3 is called the numerator and 4 is the denominator. If the numerator is smaller than the denominator, the fraction is termed "proper"; in an improper fraction the numerator is larger than the denominator, although it can be simplified to a whole number and a fraction.

word *fractus* meaning broken). In books, a fraction may be printed as ⅓ or 1/3. The number below or to the right of the line is called the denominator and is the number of parts that a unit quantity has been "broken" into. The number above or to the left of the line is the numerator and represents the number of such parts being considered. Two pieces of wood 3m and 7m long have lengths in the proportion of 3 to 7, or in the ratio 3 to 7 (often written 3:7). The shorter piece is 3/7 the length of the longer.

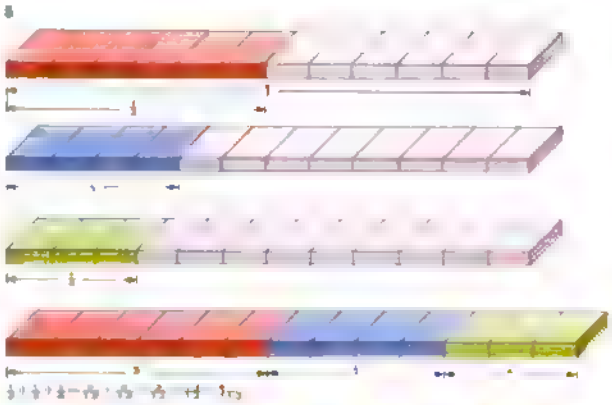
There are two types of fractions, called proper and improper. In a proper fraction, the numerator is less than the denominator: 3/7, 7/8 and 29/54 are examples. An improper fraction has a larger numerator than denominator, as in 5/4 and 22/7. Generally these are simplified by dividing out and expressing the remainder as a fraction, as in 1 1/4 and 3 1/7.

The laws of arithmetic also apply to fractions, but special techniques are sometimes needed in manipulating them. Multiplication is simple – the numerators are multiplied together and the denominators multiplied together, and the result expressed as a new fraction. Thus $2/3 \times 7/11 = 14/33$. To divide, invert the second fraction (the divisor) and multiply: $2/3 \div 7/11 = 2/3 \times 11/7 = 22/21$. Here the result is an improper fraction that can be simplified to 1 1/21 (that is, one and one-twenty first).

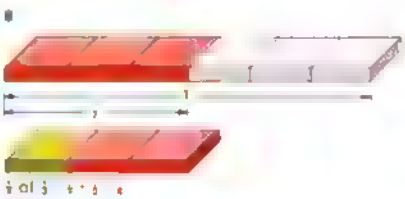
Addition and subtraction of fractions is more complicated. They must first be written in terms of the same denominator and for simplicity the smallest possible one is chosen (called the lowest common denominator or LCD). Then the numerators can be added or subtracted as necessary, the result expressed in terms of the LCD and simplified if possible [5, 6, 7].

Decimals are a way of writing fractions whose denominators are powers of ten. For example, 1 9/10, is the fraction 19/10 and is written in decimals as 1.9. The decimal point separates the whole number part (the argument) from the fractional part. Every fraction can be expressed as the sum of a series of such fractions (tenths, hundredths, thousandths, and so on) and can be represented in decimal form.

A cash register at a modern supermarket adds together the cost of each purchase. Some machines will also calculate and even issue the correct change and print a ticket.



5 Before fractions can be added they must all be expressed in terms of the same denominator. To add 1/2, 1/3 and 1/4 they must all be stated in twelfths (in this example, 12 is the lowest common denominator) as 6/12, 4/12 and 3/12. They can then be added to give 13/12, an improper fraction that simplifies to 1 1/12. This sum explains why it is impossible to divide anything into "shares" of 1/2, 1/3 and 1/4 – their sum is larger than 1.



6 To multiply fractions merely multiply the numerators and then multiply the denominators. A third of a half is $1/3 \times 1/2 = 1/6$ (the same as a half of a third – order does not matter).

7 To divide fractions invert the divisor and multiply. For example, 5/8 divided by 5 is $5/8 \times 1/5 = 1/8$, which is exactly the same as the quantity described as a third of a half, as shown in illustration 6.

8 At a public meeting a vote is often decided by "a show of hands" – those in favour of a motion raise their hands and are counted. But the way in which the results are announced or reported – as fractions or percentages – can convey different shades of meaning. At such a meeting, with 580 people present, 348 voted for the motion and 232 voted against. This basic fact can be expressed in various ways: "Three out of five people voted in favour"; "40 per cent of the voters were against the motion", and "The motion was carried with a 20 per cent majority" are all true statements based on these figures. Fractions, proportions and ratios (often expressed as percentages) are merely different ways of presenting the same information. But if 200 of the people present did not vote (abstained), a figure of "60 per cent voting in favour" means that of the 580 present only 228 people were in favour of the motion – less than half the people present.

9 Proportions are also used to define slopes – for example, gradients on roads. "A slope of one in nine" means, mathematically, that a slope rises one unit of length for every nine horizontal units. In practice distances are measured along the road's surface and a one-in-nine hill climbs one metre (or one yard) for every nine metres (yards) travelled along the road. Mathematically this hill has a slope of 1 in 8.944m – near enough for a road sign. But proportions can best be compared as percentages. Thus the ratios 7 to 13 and 28 to 53 (corresponding to fractions 7/13 and 28/53) are difficult to compare. But as the percentages 53.85% and 52.83%, the former is obviously larger than the latter.

The language of numbers

The idea of number is a basic concept. The distinction between one and many is probably the easiest for a child to understand. A boy on a beach can pick up one pebble although he can see many more. If he picks up a handful, he obviously has more than one pebble but far less than the total number he can see. To obtain a precise idea of how many he has, he can count the number of pebbles in his hand and find, for example, that there are 12. "Twelve" is the name given to that number of pebbles. It is a property possessed by all collections of 12 objects: 12 cows, 12 seagulls and 12 encyclopaedias.

Positive and negative integers

Whole numbers such as 1, 5 and 212 are called positive integers and have been used ever since men began to count. In the Middle Ages the Hindus developed the concept of negative integers to deal with amounts owing in a trading transaction. A man might own five (+5) sheep and owe three (−3), so that he really owned only 5–3 = 2 sheep.

As long as mathematical operations are limited to counting, integers are sufficient as numbers. But as soon as men started to measure they found that nature is not organized into integer lengths and areas. A farmer could make a measuring stick (a ruler) by marking off a piece of wood into similar lengths equal to, say, the length of his foot. He might find that one of his animals was 5 "feet" long, whereas its offspring was only 2 "feet" long. Then he might find an animal that was 3½ "feet" and another of 2½ "feet". He would thus discover a whole new family of numbers, called rational numbers. Any number that can be written in such a form as 8/3 – as a fraction (the ratio of two integers) – is a rational number. Such numbers can be positive or negative and all integers are rational [1A].

In the sixth century BC Greek mathematicians discovered that a square with sides one unit long has a diagonal whose length cannot be measured exactly. No matter what scale of length is used, and no matter how finely it is subdivided into fractions, such a length cannot be measured with precision nor can it be written as a fraction. The system had to be extended again to include this new class of numbers, which are now called irrational numbers [3A].

Today we use zero (0) to denote the absence of a number, but this has not always been so. The Roman numeral system, for instance, had no zero. It was introduced for its present role in about 600 BC by Hindu mathematicians who formulated rules for calculating with it: multiplying by zero always gives a zero result and addition or subtraction of zero leaves a number unaltered. Hindu mathematicians also recognized that dividing by zero does not produce a result that can be defined by the number system.

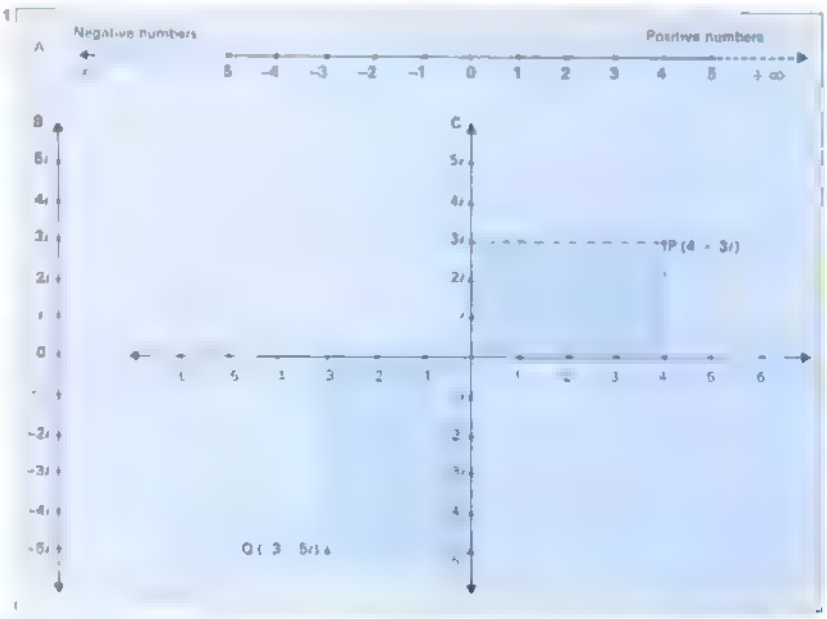
Infinite and imaginary numbers

The concept that there are infinitely large numbers was first discussed by the Greek mathematician Archimedes. Starting with the largest number in the Greek number system, "a myriad myriads" (a hundred million), he constructed even larger numbers. He then estimated the number of grains of sand in the universe and showed that this was less than his largest number.

Archimedes (*c.* 287–212 BC) showed

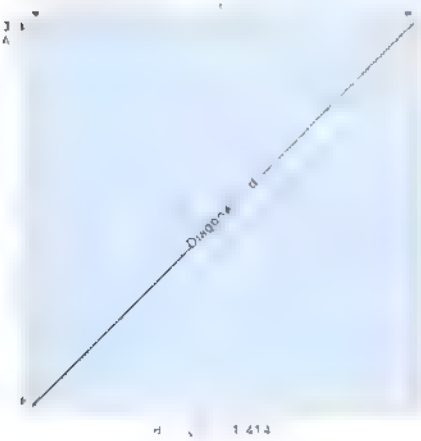
CONNECTIONS

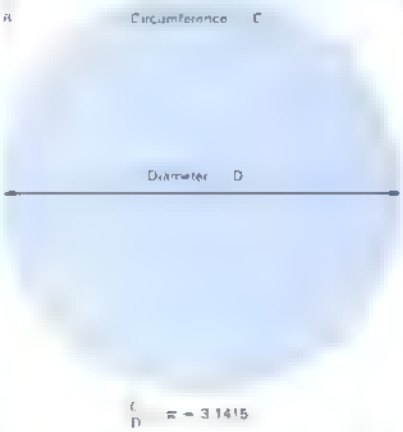
See also



1 Three types of numbers are the real, imaginary and complex numbers. Real numbers [A] can be represented as points along a line extending from minus infinity to plus infinity. They include all negative and positive numbers. Imaginary numbers [B] are based on *i*, the square root of −1, and can also be positive or negative. Complex numbers [C] each have a real and an imaginary part. They can be pictured as points defined by a distance along the real number axis and a distance along the imaginary number axis. Complex number P, for example, is 4 + 3*i*, and Q is −3 − 5*i*. Complex numbers are much used by scientists.

2 Five oranges, five hens' eggs and five bottles full of wine all possess the identical property of "fiveness". The number 5, fifth along the positive real number scale shown in illustration 1A, can be applied to any such group of five objects. The bottles do not all have the same shape, but this obviously does not affect their number. Only adding or removing some bottles would do that.





3 An irrational number cannot be expressed as a fraction using integers (whole numbers). A square with sides each one unit long [A] has a diagonal equal in length to the square root of 2. This is approximately equal to 1.414 . . . . with a never-ending series of numbers after the decimal point. Another irrational number is the ratio of the circumference of a circle to its diameter [B], represented by the Greek letter π. It is equal to 3.1415 again with a never-ending series of numbers after the decimal point. A rough approximation to π is given by the fraction (rational number) 22/7, which is equal to 3.1428. Irrational numbers were discovered by the Greek mathematician Pythagoras.

4 Keeping a tally was one of the earliest forms of counting. In this old English game called shove ha'penny the players slide coins along a board into marked off sections. They keep their scores with chalk tally marks at the edges of the board. Ancient farmers probably counted their animals using a tally-stick, a piece of wood carved with a series of notches. In some European beer halls today, the waiter gives a customer a new beer mat with each drink and keeps the old mats as a tally of the total drunk.

that there is no upper limit to a number system. Infinity, unlike zero, is not a number. No matter how large a number is there is still an indefinite number larger than it. Infinity can never be reached.

With the concepts of zero and infinity, men had a complete number system that could be pictured as every real number along a line stretching from minus infinity to plus infinity. But with the development of squares (the square of a number is that number multiplied by itself) and square roots (the square root of a number is another number which, when multiplied by itself, equals the original number), mathematicians encountered such problems as: what is the square root of −5? At first such problems were thought to be impossible to solve because there is no real number which, when squared, gives a negative result. Then in the sixteenth century Italian mathematicians introduced the "imaginary" quantity i which, when squared, gives the result −1. Numbers involving i are called imaginary numbers.

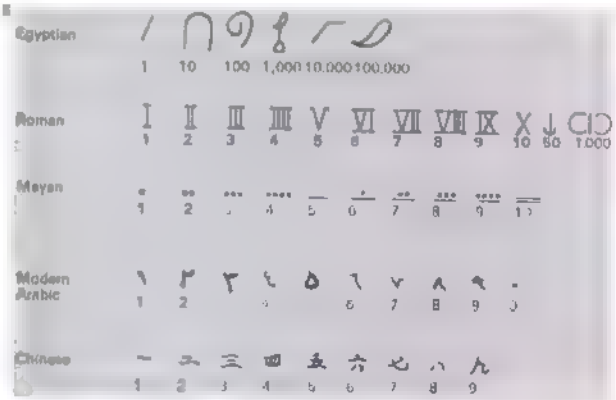
Complex numbers consist of a real part and an imaginary part, such as $5+3i$. They can be manipulated in the same way as purely real numbers. Many branches of modern engineering and electronics use them.

The system of numbers commonly used today was adapted from the Arabic numbering system [5] which, in turn, was based on Hindu ideas. In this system the position of a digit (numeral) in a number is significant. Using the basic digits 0 and 1 to 9 it is possible to construct any number. This base-10 or decimal system was introduced into Europe by Adelard of Bath in about 1100 and by 1600 was in almost universal use.

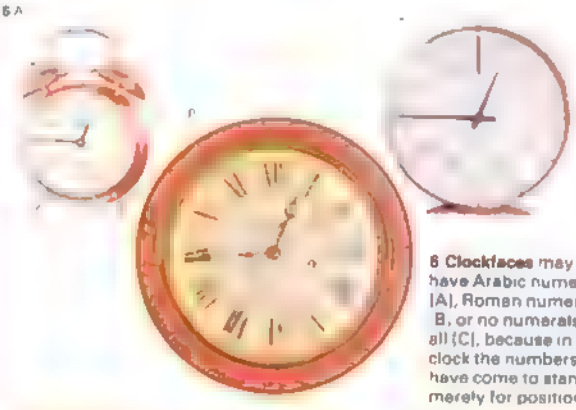
What is the base?

The base, or radix, is the number of digits in a number system. Position is important because in a number such as 333, the first 3 stands for 300 (three hundreds), the second for 30 (three tens) and the third for 3 units. But any convenient base can be used. Modern digital computers, for example, "count" using the base of 2 – the binary system of numbers – because its only digits, 1 and 0, can easily be represented by "on" or "off" pulses of electricity [9B].

An abacus is an ancient type of calculating machine still used in China and Japan. It has a number of beads on wires, generally divided into two sections with two beads (each standing for 5) and five beads (each of which stands for 1). Numbers are added or subtracted by moving the beads.



5 Various numeral systems have been used through the ages. The earliest, such as the Egyptian, used a simple pen stroke or a mark in clay to represent 1, other numbers up to 9 were formed by repeating the 1 symbol. The Romans and Mayans had an additional symbol for 5. Modern Arabic and Chinese have different symbols for each number, although 1 to 3 are formed by adding successive stokes.



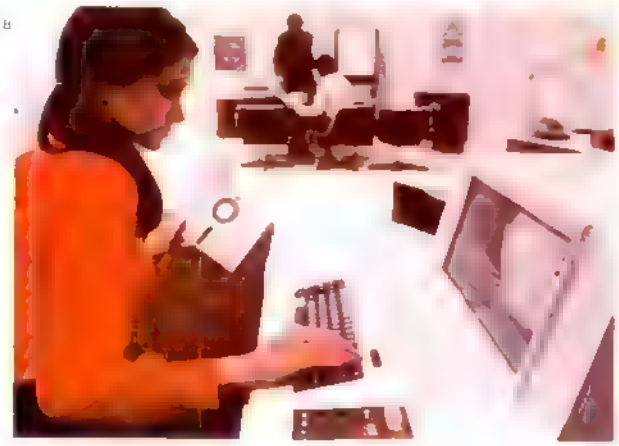
6 Clockfaces may have Arabic numerals (A), Roman numerals (B), or no numerals at all (C), because in a clock the numbers have come to stand merely for positions.

	English	French	Italian	German	Dutch	Spanish
1	One	Un	Uno	Ein	Een	Uno
2	Two	Deux	Due	Zwei	Twee	Dos
3	Three	Trois	Tre	Drei	Drie	Tres
4	Four	Quatre	Quattro	Vier	Vier	Cuatro
5	Five	Cinq	Cinque	Fünf	Vijf	Cinco
6	Six	Six	Sei	Sechs	Zes	Seis
7	Seven	Sept	Sette	Sieben	Zeven	Siete
8	Eight	Huit	Otto	Acht	Acht	Ocho
9	Nine	Neuf	Nove	Neun	Negen	Nueve
10	Ten	Dix	Dieci	Zehn	Tien	Diez

7 The names of numbers in various European languages reveal common word origins. But all these countries use the same number symbols, originally based on early Arabic numerals.

8 Large numbers are awkward to write and have different names even among English-speaking countries. American and British names are different. The scientific use of powers of 10 is unambiguous.

Number	British name	American name	Powers of ten
100	Hundred	Hundred	10^2
1,000	Thousand	Thousand	10^3
1,000,000	Million	Million	10^6
1,000,000,000	(Milliard)	Billion	10^9
1,000,000,000,000	Billion	Trillion	10^{12}
1,000,000,000,000,000	—	Quadrillion	10^{15}
1,000,000,000,000,000,000	Trillion	Quintillion	10^{18}



9 High-speed calculations are needed to deal with a changing situation in which numerical quantities vary continually, as at the tote on a race course (A) which computes winning odds for individual runners (horses or dogs), in accordance with the amount of money bet on them. For complicated calculations a computer (B) is needed in order to complete the calculation in real time – that is in time for the information to be of immediate use.

Measurement and dimensions

Four students – a chemist, a physicist, a mathematician and a humanities graduate – were each given a barometer and told to measure the height of a church tower. The chemist knew all about gases. He measured the air pressures at the top and bottom of the tower with his barometer and from the barely perceptible difference produced an answer of "anywhere between 0 and 60m" (0–200ft). The physicist was used to handling expensive equipment casually. He dropped his barometer off the tower and timed its fall, calculating the height as 27–33m (90–110ft). The mathematician compared the length of the tower's shadow with that of the barometer, arriving at a height of 30–30.5m (99–101ft). The humanities graduate sold the barometer, bought the verger a few drinks with the money, and soon found out that the tower was 30.4m (100ft) tall exactly.

Putting numbers on things

This apocryphal story illustrates the variety of ways of "putting numbers on things" and the different results that can be obtained. Life in the modern world depends greatly on man's ability to make accurate measurements, and laboratories throughout the world maintain standards of length, time, mass and voltage to ensure uniformity. Every large factory has a set of reference gauges that have been calibrated against a substandard that, in turn, has been checked against a national copy of the standard metre. As a result, a replacement bearing made in Japan can exactly fit a motor shaft made five years previously in West Germany.

Behind the practice of measurement lies theory. There is the physical theory of the process and also mathematical principles such as dimensional analysis. This derives the "dimensions" of measured quantities in terms of the fundamentals length [L], mass [M] and time [T]. Area, for example (square metres, square yards, acres or hectares), has dimensions [L^2]; volume (cubic metres, cubic yards and so on) has dimensions [L^3]. If the volume of a paraboloid were stated to be $\pi H^2/8D$, with H as its height and D its base diameter, then without making any calculations at all a student can be sure that the formula is wrong. It involves the product of two lengths divided by a length and so has dimensions [L^2/L] = [L]. It must therefore represent a length – it cannot possibly represent a volume. (The correct formula is $\pi H^2 D/8$.)

Similarly, given that the time-of-swing t of a pendulum might depend on its length l, the mass of its bob m and the acceleration imparted by gravity g, any formula for t based on a relationship between l, m and g must yield a number with the dimensions of time. Acceleration is measured in metres per second per second and has dimensions [L/T^2]. So

$$t \equiv T = \sqrt{T^2} = \sqrt{L/L/T^2} = \sqrt{l/g}$$

($\equiv$ is the sign for "is equivalent to".)

In fact the formula is $t = 2\pi\sqrt{l/g}$ (dimensional analysis can never deduce numerical factors such as 2π). Dimensional analysis is often helpful when checking calculations for errors, but one cannot derive physical laws from mathematics alone. Only by experiments can we prove that the period of swing of the pendulum is independent of its mass.

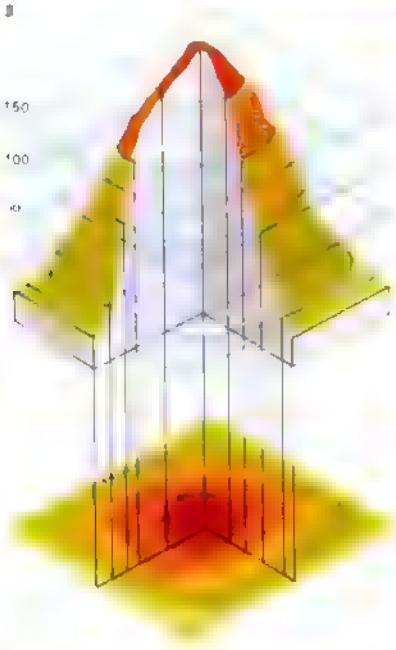
The length in metres and acceleration in metres per second per second must give an answer in seconds. "Coherent" systems of

CONNECTIONS

See also

1 Different techniques of measurement have different degrees of precision. The chart recorder (A, bottom) is precise to about 1 in 100; it is hard to read a chart more accurately than that. The chemical balance (B) can reach 1 in 10^6 and the frequency counter (A top) 1 in 10^8. Both these have a numerical display, no meter could be calibrated so finely. An air gauge (C) measures extremely small dimensions by sensing the flow of air through a small gap. Its precision of up to 1 in 10^6 is near the limits of current technology (at about 1 in 10^7 for a laser gauge). The micrometer (Key) is 10 million times less accurate.

2 Two sheets of glass in near-contact show an optical interference pattern between them. This is a contour map scaled in half wavelengths of light and makes accurate surface measurements possible.



3 Recording a three-dimensional shape on a flat surface is a problem that can be solved by the convention of contour scaling. This is shown when the cross sections of a hill at 50-, 100- and 150-metre heights are projected onto a map of the hill. The hill can be envisaged fairly well from such a map, although the "coarseness" of the contour intervals loses some finer detail. The steepness of the sides can be judged by the contour lines on the map. "Newton's rings" and similar interference patterns (as in illustration 2) are fine contour maps. Special techniques make it possible to reveal tiny deformations of stressed surfaces by means of such contours.

4 A false-colour thermogram, also known as a thermoscan, records the temperatures on the skin of a man's face. The technique allows a doctor to study the extent of skin damage caused by burns and has been adapted to aid in the diagnosis of diseases such as cancer of the breast. An infra-red camera has a rotating prism scanner that detects the heat levels in a strip of the picture, and the resulting signal is amplified and displayed on a colour television screen. Blue colours represent low temperatures, the redder the colour the "hotter" it is. Thermograms can also help architects to design houses for minimum heat loss and aid householders to insulate them.

units such as SI units (the international unit system used throughout science) guarantee an answer in correct units. The speed of a piston multiplied by its area and the pressure it exerts gives its power, for example. With units of feet per second, square inches and atmospheres, the answer would have the dimensions of power (ML^2/T^3) but be in no recognized power units. Using SI units (metres per second, square metres, and newtons per square metre) guarantees an answer in watts. Other common units are the cgs (centimetre-gramme-second) and MKS (metre-kilogramme-second) systems.

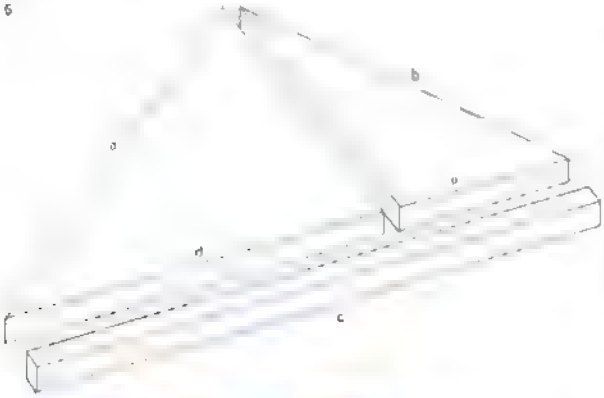
Scaling the heights

If a 10-metre (33ft) scale model of a blade of grass were actually made from grass, it would promptly collapse. Similarly a flea the size of an elephant would not be able even to stand, let alone jump. This is because the weight of an object, like its volume, increases as the cube of its height, whereas its strength increases only as its square. Many related properties scale differently so it is quite difficult, for example, to calculate what thrust will propel an aircraft from the force needed to sustain a scale model in a wind-tunnel. One way out of this difficulty is to think in terms of "dimensionless groups" like Reynold's Number, valuable in many problems of gas and fluid flow. This is LVd/η where L is a length (perhaps of a wing-section), V is a gas velocity, d is its density and η (eta) its viscosity. This combination is a dimensionless ratio – a pure number having the same value in any units.

Dimensionless ratios

Dimensionless ratios, free from arbitrary units, are fundamental entities. The ratio of the electrical to gravitational force between a proton and an electron, for example, is about 10^{39} (that is, 1 followed by 39 zeros). This is also approximately the ratio of the diameter of the knowable universe to that of the proton and of the estimated age of the universe to the time light takes to traverse a proton. The square of 10^{19}, 10^{78}, is about the number of particles in the knowable universe. Some cosmologists wonder if this ratio is trying to tell us something.

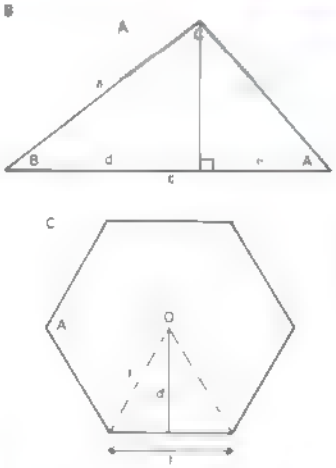
A micrometer can measure a small object with a precision of up to 1 in 10 000. This is adequate for most high-grade engineering, but far higher precision is possible.

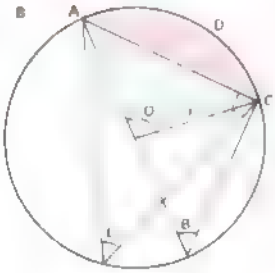


5 With e accurately known, d is found by measuring e and subtracting, error in e is diluted in the longer d. But obtaining e as c - d is inaccurate: it is a small difference between two larger numbers. Obtaining f by Pythagoras as $\sqrt{e^2 - d^2}$ is worse, squaring the two similar and large values increases their uncertainty. But e, and f are dissimilar lengths, so that b can be found accurately as $\sqrt{e^2 + f^2}$

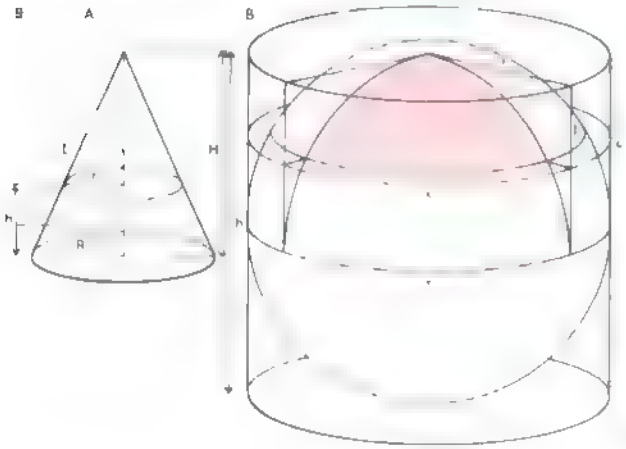
6 By the surface volume scaling law, fine structures have relatively more surface to their volume. Each die weighs 2g (0.07oz) and has 9cm² (1.4in²) of surface. The sugar lump is made of 0.5mm (0.02in) grains and has about 200cm² (31in²) of total surface. The 2g of "molecular sieve" on the watchglass is porous to the molecular level and has a remarkable 1,500m² (16,150 sq ft) of total effective surface area.

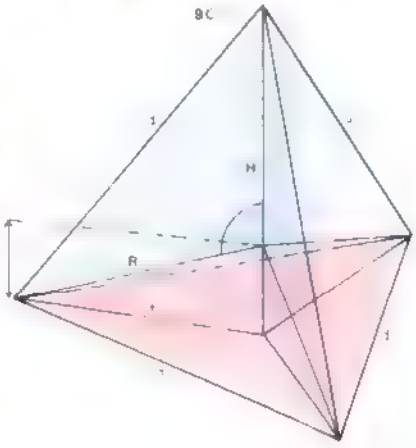
7 If the size of a balloon is doubled, its surface (and therefore weight) goes up four times, but its volume (and therefore lift) goes up eight times. This surface/volume scaling law shows that balloons become more efficient the bigger they are (A). Conversely, doubling an aircraft's size increases its weight by eight times, but the wing area by only four times (B). Small aircraft pose fewer design problems.





8 Various formulae link the dimensions of plane figures. In a triangle (A) the cosine formula states that $a^2 = b^2 + c^2 - 2bc \cos A$, and the sine formula that $a/\sin A = b/\sin B = c/\sin C$. If $S = \frac{1}{2}(a + b + c)$, the area of the triangle $= \sqrt{S(S-a)(S-b)(S-c)}$. In a circle (B), angle $O = 2B = 2E$ and the circumference of arc ADC is $\pi rO/180$. In a regular n-sided polygon (C) angle $A = 180(1 - 2/n)$, angle $O = 360/n$, the area of the polygon $= \frac{1}{2}nld$ and the radius $r = d \sec(180/n)$





9 The volumes and areas of solid figures can be linked by formulae. In a cone (A) the volume of the full cone $= \frac{1}{3}\pi R^2H$; and the area of the curved surface is given as πLR. In the frustrum of the cone (the lower section) the area of the curved surface $= \pi l(R + r)$, and the volume $= \frac{1}{3}\pi h(R^2 - Rr + r^2)$. In a sphere and cylinder (B) the surface area of the sphere $4\pi r^2$. The curved surface of the cylinder $2\pi rh$. The volume of the sphere $= 4\pi r^3/3$, and of a cylinder it is given as πr^2h. The volume of the section of sphere between planes x and y $= V_x = \pi L^2(r - L/3)$, $V_y = \pi L^2(r - L_y/3)$. In a regular tetrahedron (C) the distance $r = d/\sqrt{3}$, $H = d(\sqrt{2}/\sqrt{3})$, and $R = \frac{1}{2}d(\sqrt{3}/\sqrt{2})$. The height of the centre of gravity is given as $h = H/4$

Finding unknown quantities: algebra

In arithmetic various quantities such as lengths, areas and sums of money are represented as numbers (and the appropriate units). But some mathematical problems are concerned with *finding* a number – an unknown quantity. If two numbers add up to 10 and one of them is 6, what is the other? The answer to this simple problem is 4 and yet the method of formalizing it is a basic technique of algebra.

To solve this problem by algebra, let the unknown number be x. Then $6 + x = 10$ (this is an algebraic equation). By subtracting 6 from each side of this equation, it simplifies to $x = 10 - 6 = 4$. By making a letter, x, stand for the unknown quantity, the problem can be solved, and this way of using letters is a basic technique of algebra.

Greek and Arab mathematicians

Greek mathematicians such as Diophantus (*c.* third century AD) used letters in their equations. But the word algebra is derived from the Arabic words *al-jebr*, meaning "bone-setting" (the restoration and reduction of a bone fracture). This formed part of the title of a book by the Arab mathematician Al-Khwarizmi. By the sixteenth century mathematical problems were fully formulated in algebraic terms, initially in France by Franciscus Vieta (1540–1603). The normal convention of using the last few letters of the alphabet (x, y and z) to denote unknown quantities and the first few letters to stand for known prescribed numbers was introduced by the French mathematician René Descartes (1596–1650).

Algebraic equations and formulae

Common practical applications of algebraic equations are the various formulae used in science, particularly in mathematics and physics. The volume of a cylinder, for example, is given by the formula $V = \pi r^2 h$, where V is the volume, r is the radius of one end and h is the cylinder's height [1]. The formula provides a shorthand way of saying "the volume of a cylinder equals the area of one end multiplied by the height".

Algebraic equations [2] and formulae can be manipulated according to established rules. The subject (V) of the cylinder equation can be changed to find the radius or height of a cylinder of known volume. For instance, $h = V/\pi r^2$. Such formulae are perfectly general – they apply to all cylinders, whether they are tall and thin or short and squat. There are similar formulae for the areas and volumes of all common geometrical figures.

Many problems in algebra involve more than one unknown quantity. Consider the problem of finding two positive numbers whose product is 15 and whose difference is 2. Let the two numbers be represented by the letters x and y. Then the "product" information can be stated as the equation $xy = 15$. There are several possible solutions to this equation: 1 and 15, 3 and 5, 7.5 and 2, and so on. To proceed we must use the "difference" information, which generates the equation $y - x = 2$, rearranging to give $y = x + 2$. Substituting this expression for y in the first equation yields $x(x + 2) = 15$, or $x^2 + 2x - 15 = 0$.

Now this third equation contains only one "unknown" quantity, x. The only positive number that satisfies it is 3 (when the equa-

See also

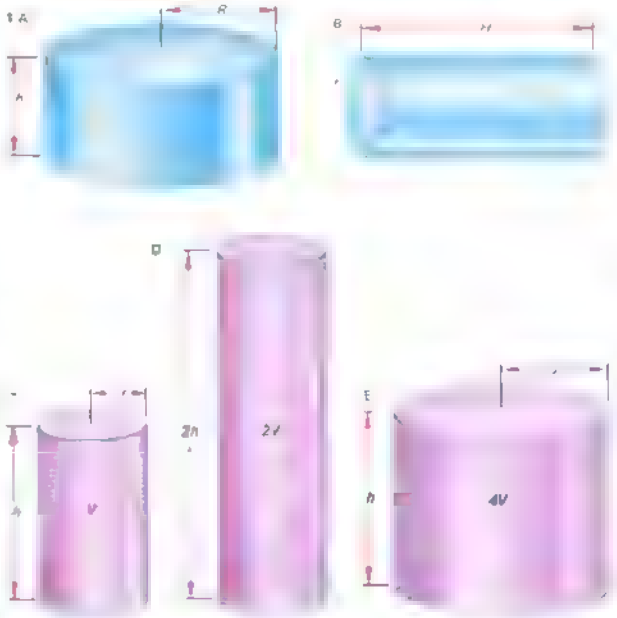
The gamma

The [illegible]

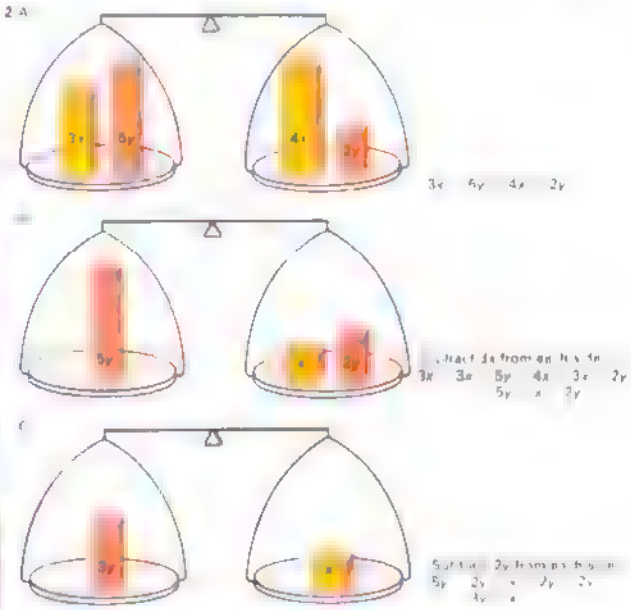
Se[illegible]

Ma[illegible]

Age of [illegible]

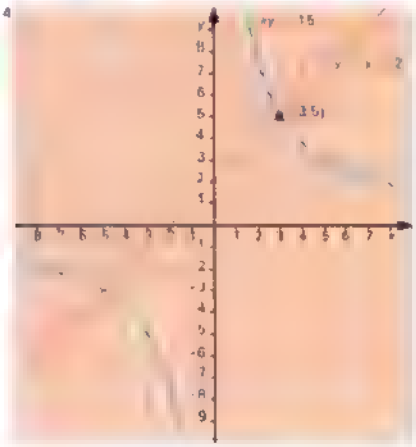


1 The formula for finding the volume of a cylinder is $V = \pi r^2 h$, where r is the radius of one end and h is the height. The two cylinders (A and B) have the same volume, but very different radii and heights. In fact, the diameter of one is almost equal to the height of the other (that is h is about equal to $2r$ and $2R$ is almost the same as H). Another cylinder (C) has volume V. Doubling its height doubles its volume (D) (for C volume = $\pi r^2 h$, for D vol. = $2\pi r^2 h$). But doubling the radius of the cylinder increases its volume four-fold (E) (for E volume is equal to $\pi(2r)^2 h = 4\pi r^2 h$). These changes are predicted using algebra.



2 An equation in algebra is in a state of balance: the terms on the left hand side taken together equal those on the right hand side, just as a collection of objects balances on a pair of scales (A). In simplifying an equation it is essential that the same operation be carried out on each side. For example, $3x$ is subtracted from each pan of the balance (B) (and from each side of the equation). A further simplification is made (C) by subtracting $2y$ from each side. As a result the original equation $3x + 5y = 4x + 2y$ reduces to $3y = x$. It is also possible to multiply or divide each side by a factor.

3 Houses in this street are numbered consecutively. A man notices that four times his number is ten more than three times his neighbour's higher up the street. What is his number? Let the house number be x, then the neighbour's is $x + 1$. So $4x = 3(x + 1) + 10$, $3x + 3 + 10$. Subtracting $3x$ from each side of this equation gives $x = 13$. The house number is 13 (and his neighbour's is number 14).



4 Algebraic equations can be plotted as lines on a graph, a technique which is the province of analytic geometry. This graph shows plots of the equations $xy = 15$ and $y = x + 2$. Treated as simultaneous equations, they are both true at the points where the lines cross. When the equation to a straight line is expressed in the form $y = mx + c$ where m and c are numbers, the letter m is a measure of the slope of the line, here equal to 1.

tion becomes $9 + 6 - 15 = 0$). Finally, to find y we substitute this value of x in either of the two original equations. According to the first, $y = 15/x = 15/3 = 5$ and for the second $y = x + 2 = 3 + 2 = 5$. The answer to the problem is therefore 3 and 5. In algebraic terms, we have solved two equations that are both true at the same time — called simultaneous equations.

By considering points in space defined by referring to their distances from a line (the x-axis) and another line (the y-axis), the equations of algebra take on a whole new meaning. The equation $xy = 15$, for example, represents a curve on which all points have the product of their x-distance and y-distance equal to 15. The equation $y = x + 2$ represents a straight line and all points along it satisfy this equation.

If these two curves are drawn [4] (to a mathematician, even a straight line is a "curve"), they intersect at the point whose x-distance is 3 and whose y-distance is 5 – the point defined as (3, 5). The graphic approach to the problem gives exactly the same solution as the purely algebraic approach. It also reveals another point at which the curves intersect, corresponding to $x = -3$ and $y = -5$. These solutions are, however, disallowed by the original problem, which called for two *positive* numbers.

The whole procedure of plotting algebraic equations as curves is the province of analytic geometry. It is the branch of mathematics in which algebra and geometry come together.

Algebra also supplies an insight into other puzzles and paradoxes. Any three-digit number whose middle digit is the sum of the other two is divisible by 11. Why? The answer can be supplied by using algebra [5].

Maintaining the balance

The examples already described serve to show the power of algebra in solving problems, particularly by manipulating equations. But there are rules about such manipulation. If there are two unknowns, such as x and y, an equation is simplified by having all the terms in x on one side and all the terms in y on the other. This can be achieved by adding or subtracting equal quantities from each side [2].

KEY

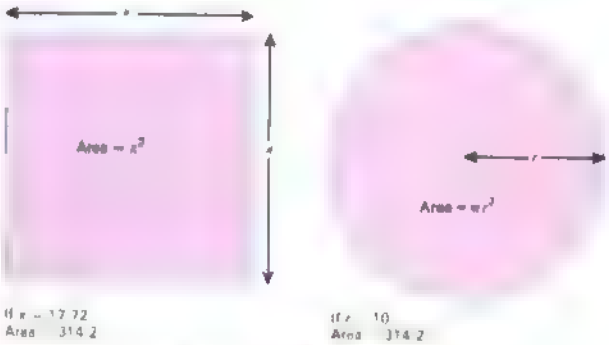
Arithmetic deals with numerical quantities — for example, the number of people in this crowd. But algebra can tackle problems involving unknown quantites, generally by allocating to them a letter such as x. Then x might stand for the number of men

5

110	220	341	473	671
121	231	352	484	682
132	242	363	495	693
143	253	374	550	770
154	264	385	561	781
165	275	396	572	792
176	286	440	583	880
187	297	451	594	891
198	330	462	660	990

5 The three-digit numbers in this table have two properties in common: the middle digit is the sum of the other two and all of them are divisible by 11. If the first digit is x and the third y, the middle one is $(x + y)$. The whole number has the value $100x + 10(x + y) + y$. This expression can be factorized and simplified to $11(10x + y)$; it is a general formula for all the numbers in the table and has 11 as a factor.

6



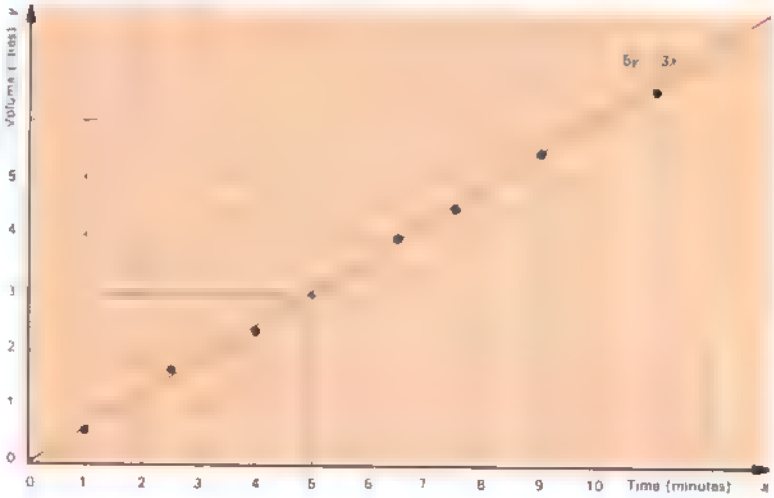
6 Squaring the circle – drawing a square with exactly the same area as a given circle – was a problem that defied ancient mathematicians. But it can be tackled using algebra. The area of a circle of radius r is πr^2 and of a square of side x is x^2. For equal areas $\pi r^2 = x^2$, or $x = r\sqrt{\pi}$. If $r = 10$, $x = \sqrt{314.2}$, which equals 17.72 approximately. A square of side 17.72cm has an area almost exactly the same as a circle of radius 10cm.

7 A

B

Minutes (x)	1.0	2.5	4.0	6.5	7.5	9.0	11.0
Litres (y)	0.6	1.7	2.3	4.0	4.5	5.5	6.6

C



7 A woman uses a photographer's clock to time the rate at which a bucket is filling with liquid [A] (the volume of the bucket is calibrated in litres on a scale inside it). The results of the measurements are shown in the table [B]. How much liquid flowed into the bucket after five minutes? To solve this problem she can draw a graph showing the rate of flow [C], which plots volume (litres) against time (minutes). The volume discharge in five minutes can be read off the graph as 3 litres. The graph is a straight line, which shows that during the time in which she made the measurement the rate of flow was constant. The line passes through the origin, at which both axes are zero. It therefore has an equation of the general form $y = mx$, where m is the slope of the line. In this example, the slope is 3/5, so the equation of the line is $y = 3x/5$ or $5y = 3x$. This equation can be used instead of the graph to calculate values of volume or time.

Mathematical curves

Anyone who can catch a ball has an intuitive grasp of mathematical curves and their transformations in space. A thrown ball travels in a mathematical curve which is (very nearly) a parabola: and many sportsmen can, with the ball still rising far away, begin to run at once to where it will fall. This is not just a simple matter of "seeing where it will go" by elementary estimation. A ball on a long piece of elastic, as is used in some tennis-trainers, is almost impossible to catch even if it is close and moving slowly. It travels in some curve other than a parabola and parabolically attuned reflexes are baffled by it

Curves, equations and laws

Gunnery fire-control and ballistic-tracking systems have to predict curves as does the sportsman. Lacking his intuitive reflexes, they need high-speed computers to represent the trajectories mathematically. This is done by deducing a precise and complete "specification" of the trajectory in the form of an equation. Mathematics is the art of making precise statements and a mathematical curve is merely one that has such a specification. It does not have to be put in equation form and indeed an informal statement is sometimes clearer Thus the statement that a circle is a curve on which every point is the same distance from a given centre is easier to understand than the curve whose mathematical equation is $x^2 + y^2 = R^2$ [2]

A mathematician can always translate the specification into an appropriate equation that is true for every point on the curve or surface but false for all other points. He can work out all the properties of the curve by manipulating algebraic symbols, which is much easier than pictorial geometry. When Thomas Telford (1757–1834) built the suspension bridge over the Menai Strait, Wales, in 1826, he had to determine the curve of the hanging chains by setting up a large model across a dry valley and measuring it – a sad consequence of mathematical ignorance. Nowadays an engineer can derive the equation of a suspension bridge cable and find out all he needs to know about it without even drawing a diagram

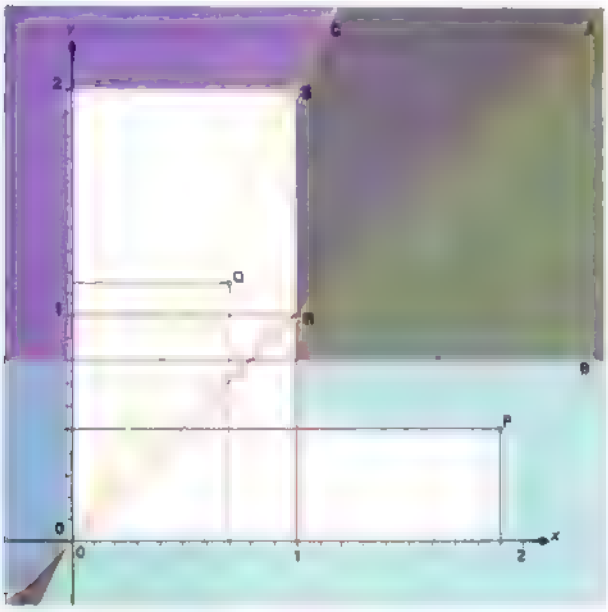
Because the world is governed by simple mathematical laws, mathematical curves are all around us. A stone falls in a straight line if dropped or a parabola if thrown; the Moon and artificial satellites move in (very nearly) ellipses, the Sun and Earth are nearly spherical, and a static liquid surface is (very nearly) flat – all because of the mathematical form of the law of gravity. A rainbow is a circular arc and the bright cusp you sometimes see in a sun-illuminated cup or pan is an epicycloid [9B] because of the laws of optics. Indeed much of the process of scientific discovery consists of observing such things, or conducting experiments to reveal them and then deducing "laws" that must hold to give rise to them. But most often the scientist finds mathematical curves in his results only when he draws them (plots them as a graph). Then the equation of the graph tells him the "law" revealed by his experiment

Spheres occur in nature in objects that adopt this shape as the "line of least resistance" to forces affecting them. Small droplets of water and soap bubbles are spherical to minimize their areas under the effect of surface tension. Lead shot was once made by pouring molten lead down inside a tall

CONNECTIONS

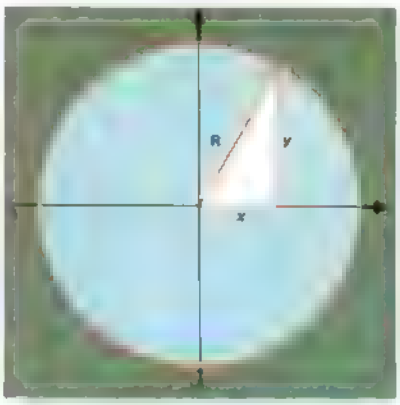
See also

1



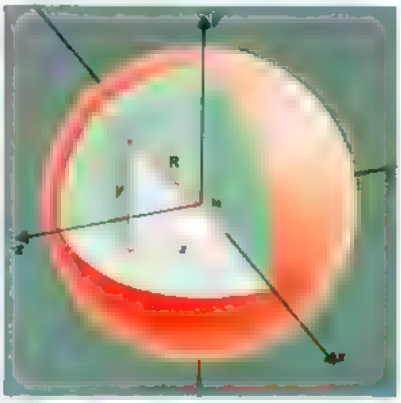
1 To give a mathematical curve an equation, first draw two lines at right angles the x axis ([0x] on the diagram) and the y axis [0y] Then every point on the paper can be defined by its "x distance" and "y distance" along these axes. Thus point Q has $x = 0.7$ and $y = 1.15$, point P has $x = 1.9$ and $y = 0.5$. On the straight line A it is obvious that for every point on it (eg R, with $x = 1$ $y = 1$) the x distance equals the y distance or $x = y$. This then is the equation of line A as a mathematical curve, true for all points on it and false for all others. Line C has equation $y = 2x$ (as point S shows), line B, equation $y = 0.8$.

2



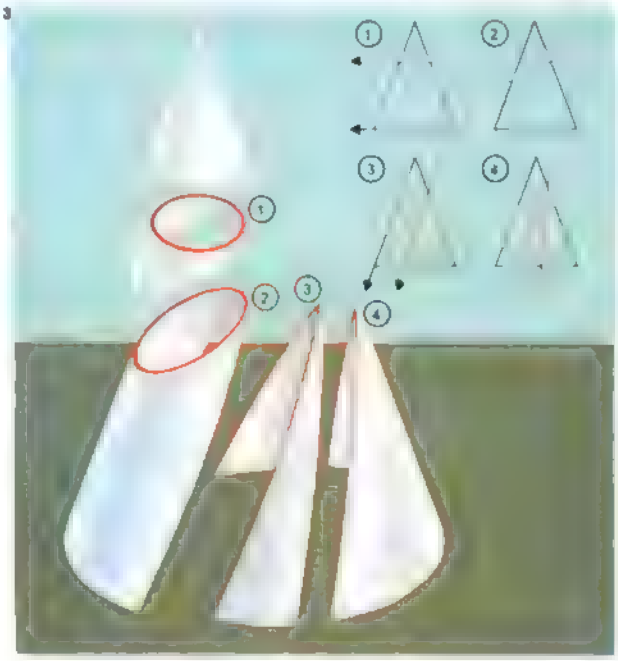
2 To find the equation of a circle, observe that each point on it forms a right-angled triangle whose sides are the x distance and y distance and whose hypotenuse is R. Pythagoras's theorem tells us that $x^2 + y^2 = R^2$ For points outside the circle $x^2 + y^2$ exceeds R^2 inside it $x^2 + y^2$ is less than R^2 Along the intersection of these two regions the equation balances All mathematical curves divide space in this way

4



4 A mathematical surface can be defined like a curve, using three mutually perpendicular axes in space, x, y, z. The equation of the sphere is $x^2+y^2+z^2 = R^2$ All points outside it have $x^2 + y^2 + z^2$ greater than R^2, those inside have $x^2+y^2+z^2$ less than R^2, the equation balances on the boundary An equation such as $x^2+y^2+z^2 - 2x - 8z+17 = R^2$ also defines a sphere but its centre is not at the intersection of the three axes

3



3 An important family of mathematics curves results from sectioning a cone at various angles A horizontal section [1] gives a circle an inclined one [2] an ellipse. A section parallel to one side of the cone [3] gives a parabola and still greater inclination [4] a hyperbola All these have the same general equation $ax^2 - by^2 + 2hxy + 2gx + 2fy = c$ With h^2 greater than ab, it is a hyperbola with $h^2 = ab$, a parabola, h^2 less than ab gives an ellipse, of which the circle ($h = 0$, $a = b$) is a special case The terms a, b, c, h, g and f are chosen constants, with $b = c = h = g = 0$, $a = 1$ and $2f = 1$, the equation becomes $y = x^2$ As $h^2 = ab$ (both $= 0$) this is the equation of a parabola

5

5 A parabola rotated about its axis of symmetry gives a mathematical surface, the paraboloid. A uniformly rotating liquid acquires a paraboloidal surface from the interaction of gravity and centrifugal force, as can be seen by spinning a pan of liquid on a turntable. This surface is perfect for radio- and optical telescope mirrors a mirror surface can be made by spinning plastic resin in a dish while it sets and coating it with metal

building into a tank of cold water. During their fall down the tower the droplets of lead formed spheres and cooled sufficiently to retain their shape when they hit the water

From theory into practice

Mathematical curves and surfaces are involved in all sorts of human activities. Lenses, for example, have spherical surfaces not because this is the ideal optical shape (although it is quite good) but because it is so easy to make. A spherical surface is the only one that presents the same form and curvature no matter how it is turned around. Therefore a hard surface and a soft one, rubbed together with back-and-forth and mutual twisting movements, will wear together into mating spherical surfaces as these are the only ones that enable them to fit exactly together every time. As a result, simple grinding processes suffice to produce spherical surfaces; where other surfaces are needed (as in telescope corrector-plates and some zoom lenses), manufacture is much more difficult.

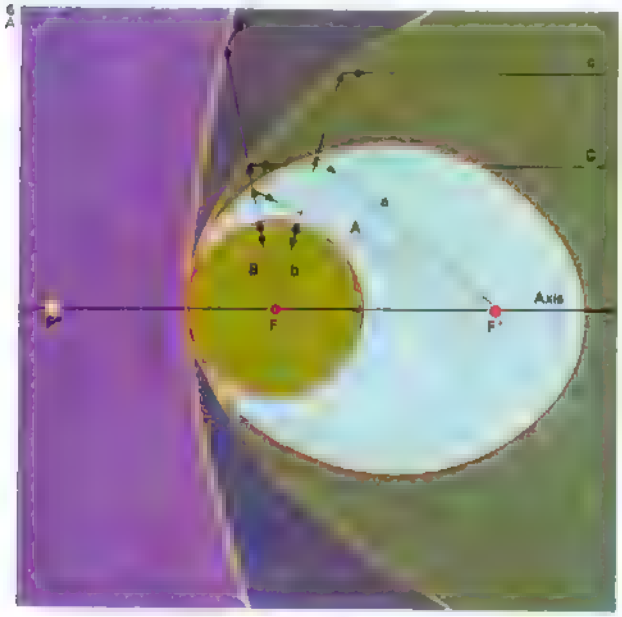
Similarly, cylindrical objects (tubes, rods, bolts) and holes are so common because this type of surface is readily created by rotating machinery. Boilers and pressure vessels are cylindrical because this shape resists pressure better than others. Cooling towers are hyperboloidal because the two-way "saddle" curvature resists weight and wind loadings well. A trumpet has an exponentially flared horn because this is the mathematically ideal way of launching the intense sound vibrations in its throat into the open atmosphere.

Beauty and the mathematician

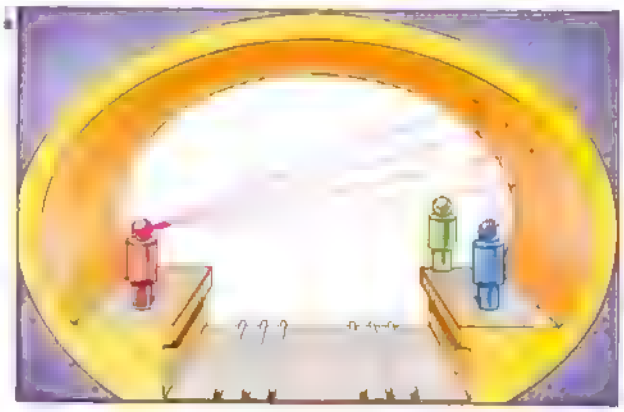
The cooling tower, trumpet, bridge arch and radio-telescope dish all derive their forms from pure mathematical physics. Yet they, and other engineering creations, have an aesthetic appeal that is sometimes sadly missing in those architectural and automobile creations whose "styling" lacks any mathematical necessity. It is an appealing flight of fancy to wonder whether the same intuition that enables the ordinary man to recognize and respond to the unity and inevitability of mathematical curves at work, also enables him to catch a moving ball.

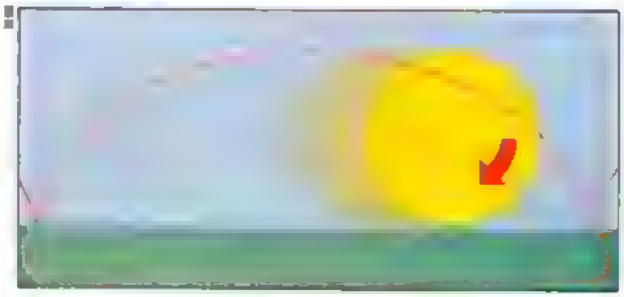
KEY

These three bridges spanning the River Tyne, are curves at work. The arches are close to parabolas of different forms. The parabola is ideal for an arch whose weight is negligible compared to the even weight of the roadway. For an actual bridge the arch's own weight must be allowed for.

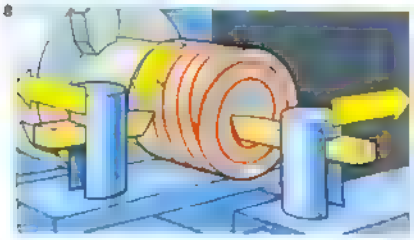


6 The conic sections (A) all have characteristic "reflecting" properties. An ellipse has two foci (F, F') such that rays emitted from one converge (A, a) on the other. A circle is an ellipse with its foci coincident and an emitted ray is reflected straight back (B, b). A parabola is like an ellipse with the other focus infinitely far away: rays emitted from the true focus are reflected parallel to its true axis (C, c). The hyperbola reflects rays from F as if they came from another focus F' behind it. Paris Metro tunnels are almost elliptical and whispering on one platform can be heard on the other by the focusing effect (B).



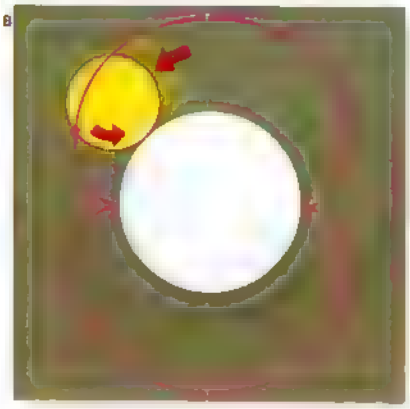


7 Many living creatures grow by compound interest, the bigger they become the faster they grow. The pearly nautilus of the Indian and Pacific oceans extends its shell continuously in spiral fashion as it grows, generating a natural "logarithmic spiral." This mathematical curve is also called the "equi-angular spiral" because a line drawn outwards from the centre intersects the curve at the same angle.



8 An "Archimedean spiral" is traced by a point that travels around a centre, varying in distance from it in proportion to the angle it moves through. Similarly a helix is the curve of a point that moves around a cylinder, travelling along it in proportion to its total angular rotation. Both are created automatically by a lathe whose cutting tool traverses a rotating workpiece. Many cylindrical objects show their machine finishing as a fine helical pattern.

9 A point on the rim of a rolling wheel traces out a cycloid (A). This famous mathematical curve received attention from Galileo, who suggested it as an arch form for stone bridges, and Newton, who proved that inverted it was that surface down which a particle slid in the minimum time. A wheel rolling on another wheel generates an epicycloid (B). To mesh smoothly without "stepping" from tooth to tooth, gear wheels need teeth of defined curves. The rise of an epicycloid from the inner wheel is one ideal curve. The point on the rim of a wheel rolling inside another wheel generates a curve called a hypocycloid, also used in gears.



Logarithms and slide rule

As the science of mathematics and its applications progressed, men found themselves having to carry out more and more complicated calculations – especially ones involving multiplication and division. Even a modern computer or electronic calculator takes about ten times as long to multiply two numbers as it does to add them, and this is certainly true of human mathematicians

Both multiplication and division were simplified in the 1500s by the introduction of decimal notation. Then a Scottish mathematician, John Napier (1550–1617), published his book *Mirifici Logarithmorum Canonis Descriptio* (1614) which announced the discovery of logarithms

Arithmetic and geometric progressions

A practical form of Napier's ideas about logarithms was a set of numbered rods, or bones, that could be used for carrying out multiplication by merely using the mathematical operation of addition [1]. As with logarithms (and the slide rule) they made use of two types of mathematical series, called arithmetic and geometric

An arithmetic progression is a series of numbers in which each is obtained by adding a "common difference" to the one before it in the series. The ordinary ordinal sequence of numbers, for example – 1, 2, 3, 4, and so on is an arithmetic series with a common difference of 1. In a geometric progression, each term is obtained from the previous one by multiplying it by a "common ratio". In the series 2, 4, 8, 16, and so on, the common ratio is 2

In the three following series

1	2	3	4	5
10	100	1,000	10,000	100,000
10^1	10^2	10^3	10^4	10^5 .

the first is an arithmetic progression and the second a geometric one (with a common ratio of 10). The third row, equivalent to the second, shows how the succeeding powers of 10 in the second (geometric) series are in an arithmetic progression

The powers (exponents) in the bottom row are called the logarithms of the corresponding terms in the middle row "to the base 10" The logarithm of a given number is the power to which a fixed number (the base) must be raised in order to equal the given number Thus the logarithm of 100 to the base 10 is 2 (because $10^2 = 100$)

Logarithms, pianos and guitars

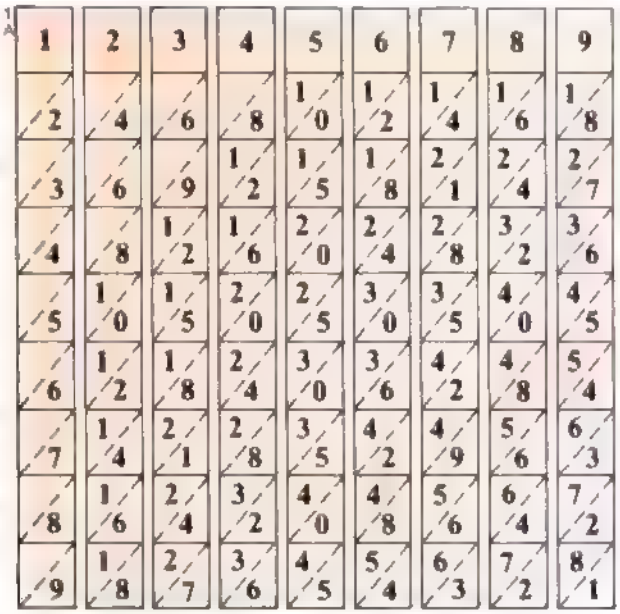
Ten is not the only base for logarithms. Napier's original tables were to the base "e" (an irrational number) and are still much used in science, where they are called natural or Naperian logarithms. The pitch of the notes on a piano is in a logarithmic ratio to the base 2, whereas the keys are in a linear sequence of octaves

The sound wavelength of any note is twice that of the one an octave above it The pieces of metal called frets across the fingerboard of a guitar also form a logarithmic series (in terms of spacing)

Numbers expressed in terms of powers or exponents are multiplied by adding the exponents. Thus $10^2 \times 10^4 = 10^{2+4} = 10^6$ And since logarithms are also exponents, to multiply two numbers their logarithms are merely added and tables can supply the number whose logarithm is the result. In this way multiplication is reduced to the much easier

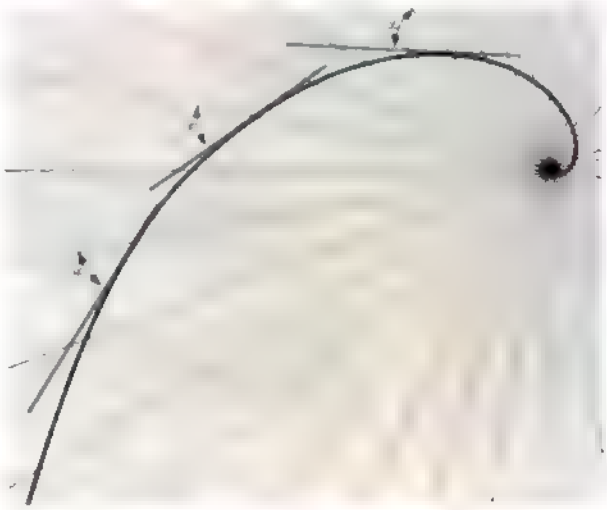
CONNECTIONS

See also



1 Napier's bones consisted of a set of nine square section rods (A), housed in a tray They were numbered 1 to 9 in the first segment and the lower segments on each rod were divided diagonally These segments were numbered down the rods in arithmetic series, the "one" rod numbers increasing by 1 (to give 1, 2, 3, 4, and so on), the "two" rod by 2 (2, 4, 6, etc), the "three" rod by 3 (3, 6, 9, etc) and so on for the whole set ending with the "nine" rod which was calibrated in nines (9, 18, 27, 36, etc) The other faces of each square rod were similarly calibrated, so that each number (1 to 9) had to be represented four times somewhere in the set To find the multiples of a number – for example 1,572 the rods numbered 1 5 7 and 2 are removed from the tray and laid side by side (B). To find 3 × 1,572 the third row of rod segments is used, as at C The numbers displayed can be added diagonally as shown, to yield 4,716, which is the required product. To find 8 × 1,572 the segments in the eighth row are used, as at D When added diagonally, the numbers displayed this time add to 12,576 – again the required product. To multiply by a larger number say 38, the appropriate products are merely added together (47,160 – a zero is added because we are now multiplying by 30 not 3 – and 12,576) to give 59,736

2 The frequency of a note in music is twice that of the one an octave below it On a keyboard instrument the frequencies of a note and its successive octaves are in the proportions 1 2 4 8 16, etc This is a logarithmic scale to the base 2 The spacing of the metal frets across the fingerboard of a guitar are also in a logarithmic sequence and by pressing his fingers against each in turn the guitarist is able to play the notes up a chromatic scale. An unfretted instrument such as a violin works on the same principle but the divisions are not marked on the fingerboard



3 In this curve the angle α between a tangent at any point and the radius drawn from the centre is constant. For this reason It is called an equiangular spiral. The lengths of the radii to the curve are proportional to the logarithms of the angles between the radii and the initial horizontal direction, it is therefore also known as a logarithmic spiral

task of addition. Similarly, logarithms can be used to perform division by actually carrying out a subtraction.

To calculate in decimal numbers, logarithm tables need be compiled only for the numbers between 0 and 9 999 (in four-figure tables; five-figure tables include 9 9999, and so on to as many figures as required). Larger numbers are expressed by adding a whole number (integer) called the characteristic, which represents in base-10 logs the corresponding power. The four-figure logarithm to the base 10 (written $\log_{10}$) of 2, for example, is 0.3010 [4]. The log of 200 is 2 3010 and of 2,000 is 3.3010 (200 is $10^2 \times 10^{0.3010} = 10^{2.3010}$, and 2,000 is $10^3 \times 10^{0.3010} = 10^{3.3010}$).

The slide rule

A slide rule [5] is a mechanical device for multiplying and dividing numbers to limited accuracy. Logarithmic scales are engraved on rods that can be slid in relation to each other, and numbers on them added or subtracted as needed – added for multiplication and subtracted for division. Because of the log scale, the numbers become closer together along the slide, just like the frets on a guitar's fingerboard. Unlike an ordinary ruler the scale is geometric rather than arithmetic [6]

In its simplest form a slide rule has only two scales – called the *X* or *D* scales on a complicated slide rule. To multiply two numbers, the 1 on the upper scale is set opposite one of these numbers on the lower scale and the required product read off it opposite the second number on the upper scale [6A]. For division the two numbers are lined up and the quotient read off against the 1 [6B]. A transparent sliding "saddle" called a cursor can be moved along to line up the graduations and make them easier to read

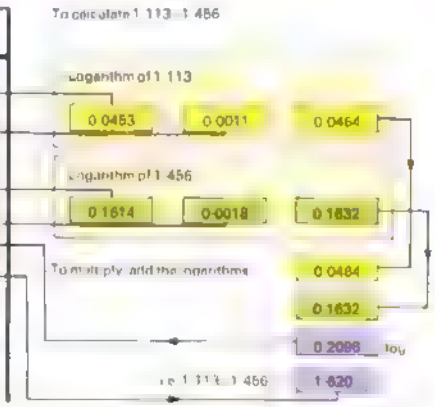
The accuracy of a slide rule is limited mainly by its length. A cylindrical slide rule [7] has scales up to a metre long wound round it like a screw thread. Most ordinary slide rules have additional scales to aid various types of calculations, reciprocal (a scale of all the numbers divided into 1), square (numbers multiplied by themselves), square root and even trigonometrical functions such as sine, cosine, and tangent

KEY

The curve shown in illustration 3 is a logarithmic spiral Such curves occur in nature, generally revealing the effects of accelerating growth as in the spiral shells of snails and various other molluscs and in flowers like this one

4

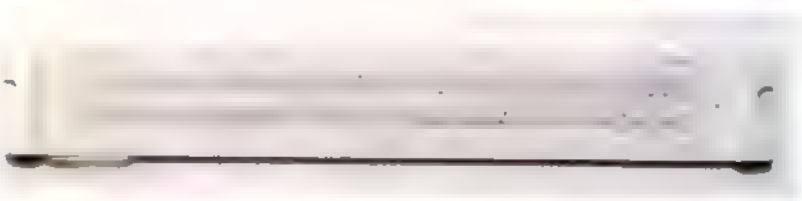
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19



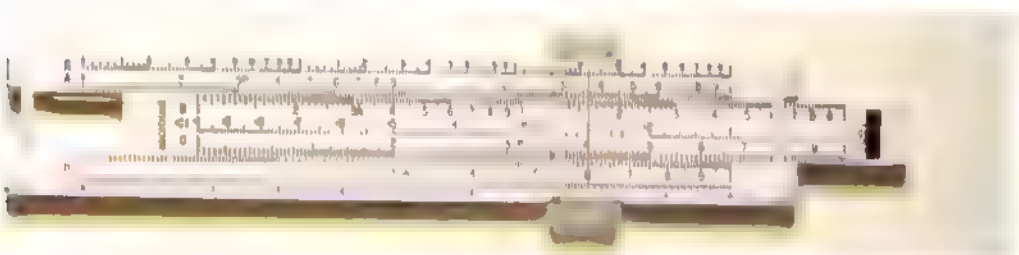
4 Log tables can be used to multiply or divide numbers. In this example 1 113 is found by adding the log under 1.11 to that for 0.003, to give 0.0464. Similarly the log of 1.456 is 0.1632 The two logs are then added to give 0.2096 and the required product is the number that has this logarithm – the number 1 620 in the table In practice, results are found by consulting tables of antilogarithms

5 A modern slide rule has various scales such as x, x^2, $1/x$, root x and so on Some also have trigonometrical and other functions for calculations by navigators, engineers and others who use them

5



6
A



B

6 To multiply using a slide rule (A), for example to find the product of 1.5 and 4, the 1 on the upper scale is lined up with 1.5 on the lower scale. The required product is read off on the lower scale opposite the 4 on the upper – in this example to give 6. The example of division (B) is 6÷3. The 3 on the upper scale is lined up with the 6 on the lower scale and the answer read off opposite the 1 on the upper scale. In this example the required answer is 2

7

7 A cylindrical slide rule can be pictured as a set of long scales wound round the cylinder like a screw thread

Sets and groups

The mathematical theory of sets was first investigated by Georg Cantor (1845–1918) and later systematized by Ernst Zermelo (1871–1956), but the basic concepts were known earlier. Some adults new to these ideas find them difficult, but children have an intuitive grasp of them early in life. The concepts of number and operations on numbers are abstractions from the experience of sorting and combining sets of objects.

Collecting objects together

The idea of a set is the most fundamental concept in mathematics. A set is a collection of objects with a common description or definition, listed in any order or according to a formal law. The set of oceans, for example, is defined as: oceans = {Pacific, Atlantic, Indian, Arctic, Antarctic} or O = {x | x is an ocean}. The letter O labels the set; x is called a variable; { and } are called braces, and the symbol | means "where" or "such that". This kind of set is a finite set because its cardinality (number of elements) is finite – it has a known value, in this case five. The set of counting numbers is an infinite set because we cannot say exactly how many elements it has: counting numbers = {1, 2, 3, ...}, or C = {x | x is a counting number}.

The set of natural numbers is J^+ = {1, 2, 3, }, with the same elements as the set of counting numbers. We say that C and J^+ are equal sets. Sets with the same cardinality are called equivalent sets: the set {blue, green, yellow, orange, red} is equivalent to the set of oceans – they each have five elements.

The language of sets can be understood by studying a particular example. A universal set [1], the set of all elements under consideration, can be partitioned into what are called disjoint subsets – that is, non-overlapping sets. If there are only two such sets, one is the complement of the other [2]. The set of elephants living at the North Pole is an example of the empty or null set, since it has no elements. The null set is written as ∅. In illustration 2, for example, there is no intersection of sets A and B, or P and C, so the intersection equals ∅. The concepts of partition, complement, intersection [3] and union [4] are fundamental to the processes of classification of information.

Networks [5] give rise to the Cartesian product of two sets. This is obtained by finding all possible ordered pairs of elements, taking one from each set. The word Cartesian is derived from René Descartes (1596–1650), who propounded the concept of co-ordinates. If set X is associated with the infinite set of points making one line in a plane and set Y is associated with the infinite set of points making another intersecting line, the Cartesian product of X and Y is associated with the infinite set of points making the plane containing the two lines [6].

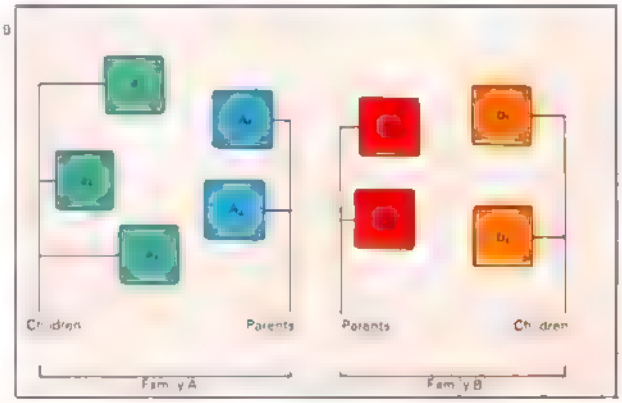
Boolean and propositional algebra

The algebra of sets is known as Boolean algebra. It is isomorphic – that is, it has a one-to-one correspondence – with the algebra of propositions or logic. It is named after George Boole (1815–64), who founded the modern study of logic. The two types of algebra use different symbols, with *union* (∪) and *intersection* (∩) corresponding to *or* (∨) and *and* (∧). Propositional algebra analyses the sets of logical possibilities in which various statements and combinations

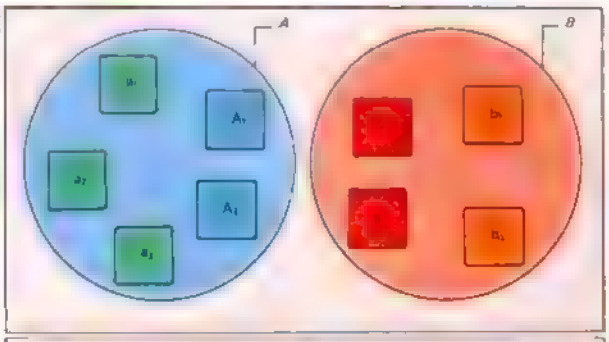
CONNECTIONS

See also

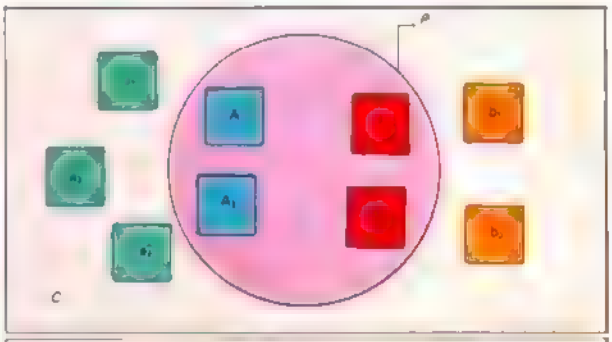
1 Two families – one with three children and one with two – together make up a universal set (A). It can be represented diagrammatically (B) and letters given to each of the elements in the set. The letters are then sufficient for the mathematical manipulation of the set in what are called Venn diagrams, first introduced by the mathematician John Venn (1834–1923) in 1880. In such diagrams, areas represent sets of things.



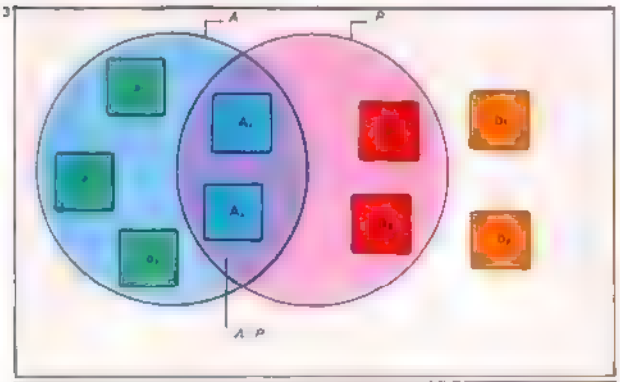
2 These Venn diagrams show how the universal set of illustration (1B) can be split into two non-overlapping subsets. Each family can make up a subset (A) or the parents and children can each form subsets (B). In each case the subsets are complementary to each other because they include between them all the elements of the first universal set. The complementary relationships in (A), for example, are written as $A - B$ and $B - A$.



Complementary sets $A - B$, $B - A$



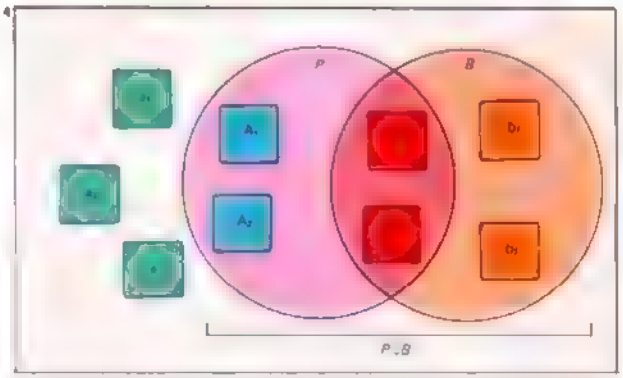
Complementary sets $P - C$, $C - P$



$A \cap P$, $P \cap A = \{A_1, A_2\}$

3 Intersection of sets generates another subset that contains all the elements common to both. Here the intersection of A and P (written as $A \cap P$) gives a subset containing only A_1 and A_2.

4 Union of sets generates yet another subset that contains all the elements in two original sets, in this example B and P. It is written $B \cup P$ and contains the elements A_1, A_2, B_1, B_2, b_1 and b_2. $B \cup P = P \cup B$ illustrates the commutative law.



$B \cup P$, $P \cup B$ $\{A_1, A_2, B_1, B_2, b_1, b_2\}$

of the statements are either true or false.

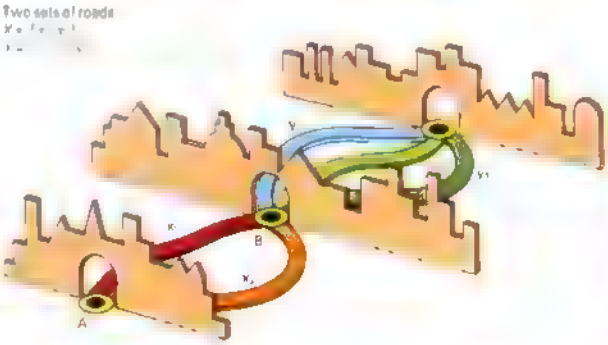
A mathematical system is created when one or more binary operations are applied to a set of elements. A binary operation combines two elements into a third of the same set. One of the most valuable systems is the "group", as it occurs in many diverse situations and helps to unify the study of mathematics. The theory was developed by Evariste Galois (1811–32) and later systematized by Arthur Cayley (1821–95). The concept of a group can be illustrated by studying a simple case of formation dancing [8] in which four dancers change their positions (or remain still) to form various patterns. The movements form the set and the operation is a combination of movements called "follows", indicated mathematically by the symbol $\otimes$. Combining any two movements results in one of the four. The identity element is I and each element is its own inverse in this particular example. We have the relationships $(J \otimes K) \otimes L = L \otimes L = I$ and $J \otimes (K \otimes L) = J \otimes J = I$, so that the associative law $(J \otimes K) \otimes L = J \otimes (K \otimes L)$ is valid. In the particular example of dancing there is another law, the commutative law $A \otimes B = B \otimes A$.

From the four possible choices available in moving a rectangle [9], a set of four transformations arises. These can be paired by the operation "follows" to produce a combination of movements that is in a one-to-one correspondence to those in the dancing example. The two types are said to be isomorphic. The search for isomorphisms is essentially the core of mathematical study.

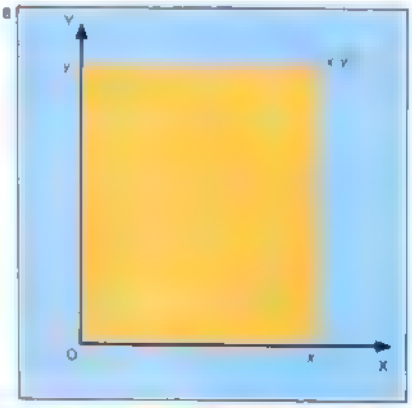
The usefulness of group theory

Group theory is useful in the study of number systems. The set of integers (whole numbers) with 0 included, $\{\ldots -3, -2, -1, 0, +1, +2, +3, \ldots\}$, is a group under addition with 0 as the identity element. The set of rational numbers is a group under addition if 0 is included. It is a group under multiplication if 0 is excluded. The use of group theory in the study of arithmetic not only enriches it but leads to higher concepts such as those of rings and field, sets of elements subject to two binary operations (addition and multiplication) that satisfy certain axioms.

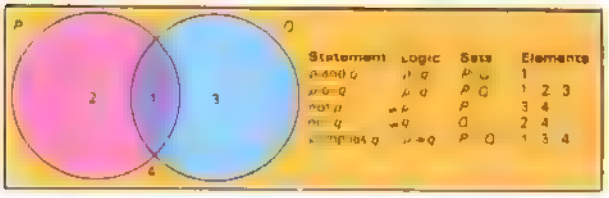
Any collection of objects constitutes a mathematical set - a collection of cans of soup, bottles of vinegar, supermarket trolleys or people in the shop are all definable sets



5 A map shows the roads connecting two towns A and C. All the roads pass through town B. The two routes from A to B are one set and the three between B and C another set. There are six possible ways of going between A and C. This is known as the Cartesian product of two sets, in this case all possible combinations of paired elements taking one from each set. The study of networks is one aspect of the subject of topology.

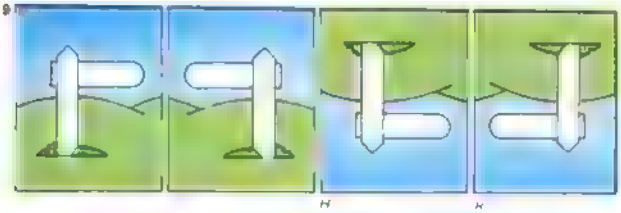


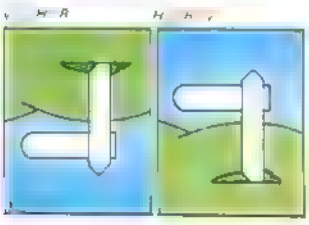
6 A plane defined by two lines is related to the Cartesian product of two sets that represent an infinite number of points on the lines. The point at the corner of the plane is defined by the co-ordinates x and y, written (x,y). These are called Cartesian co-ordinates and are used in co-ordinate or analytic geometry in which all lines whether straight or curved, can be expressed in terms of algebraic equations.



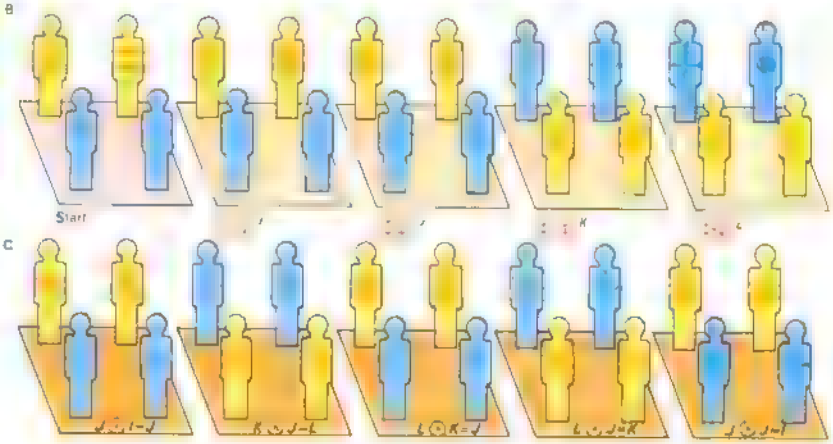
7 Union and intersection in set theory correspond to "or" and "and" in logic. This relationship enables particular elements and combinations of elements from sets to be defined by logical statements.

8 Four dancers [A], starting at the corners of a square, can have various positions [B], represented by the symbols I, J, K and L. Carrying out pairs of movements, one after the other, results in new position [C] described as J follows I, for example. Sequences of three movements can be analysed as two and the final position predicted. L follows K follows J, for example, reduces to L follows L, equals I.





9 Symmetries of a rectangle involve rotating it in various ways – vertically V, horizontally H or in the plane of the picture R. The letter I represents its original position. Again successive pairs of movements always result in one of the four.



Finding changing quantities: calculus

In a "political vocabularly" compiled for the daily newspaper the *Guardian*, the noun "decrease" was cynically defined as: "reduction in rate of increase – as in unemployment, crime, inflation, taxation, etc". This not only exposes official double-talk [2] but also highlights the universality of a central concept of calculus – rate of change

Rates of change became important in physics in 1638 when Galileo (1564–1642) concluded that a falling or thrown body had a downward velocity that increases steadily – that is, its rate of increase of downward velocity is constant [1]. What then is its trajectory? It took the genius of Isaac Newton (1642–1727) and Gottfried Leibniz (1646–1716) to solve this problem neatly and completely; the tool they created for the job was calculus.

Velocity, said Newton, is rate of change of position with time – 60km/h, for instance Similarly, acceleration is rate of change of velocity with time A car that takes three seconds to reach a speed of 60km/h from a standing start has an average acceleration of 20km/h per second. Galileo's law is that the downward acceleration of a falling body is constant. Calculus provides the methods of obtaining velocity from acceleration and position from velocity, and the whole problem is neatly solved. The calculus operation of deriving velocity from position, for example, is called differentiation Its inverse is called integration

The simplicity in symbolism

All mathematics is a sort of symbolic machinery for making subtle conceptual deductions without having to think them out – all the thought has been built into symbolism. Long division, for example, enables 431,613 to be divided by 357 by the unthinking application of a few rules; it is not necessary to know why it works, or what division really means. Calculus is perhaps the supreme example of a symbolism whose economical elegance reduces intractably complex and elusive problems to back-of-an-envelope simplicity

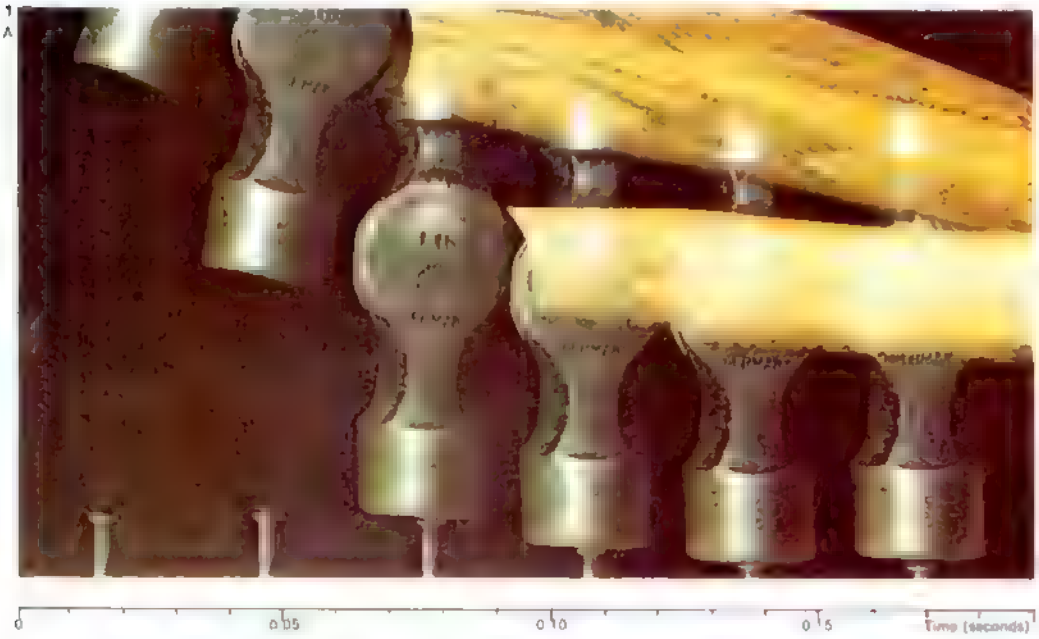
In mechanics, the branch of physics for which calculus was invented, it is omnipresent in Newton's second law of motion force equals mass multiplied by acceleration Given any two of these quantities, the equation defines the third Consider an internal combustion engine. What is the instantaneous acceleration of the piston as the crank passes top dead centre? Calculus provides the answer so that, knowing the piston's mass, it is possible to find the force on it that must be withstood by the connecting-rod For what speed will this force become excessive? Again, the pressure on the piston during the power stroke is changing every instant with the burning of the charge and the changing volume of the gases in the cylinder as the piston descends What then is the total energy imparted to the piston by the whole stroke and for what moment of ignition is it a maximum? This and a myriad of other mechanical problems could hardly be formulated, let alone solved, without calculus

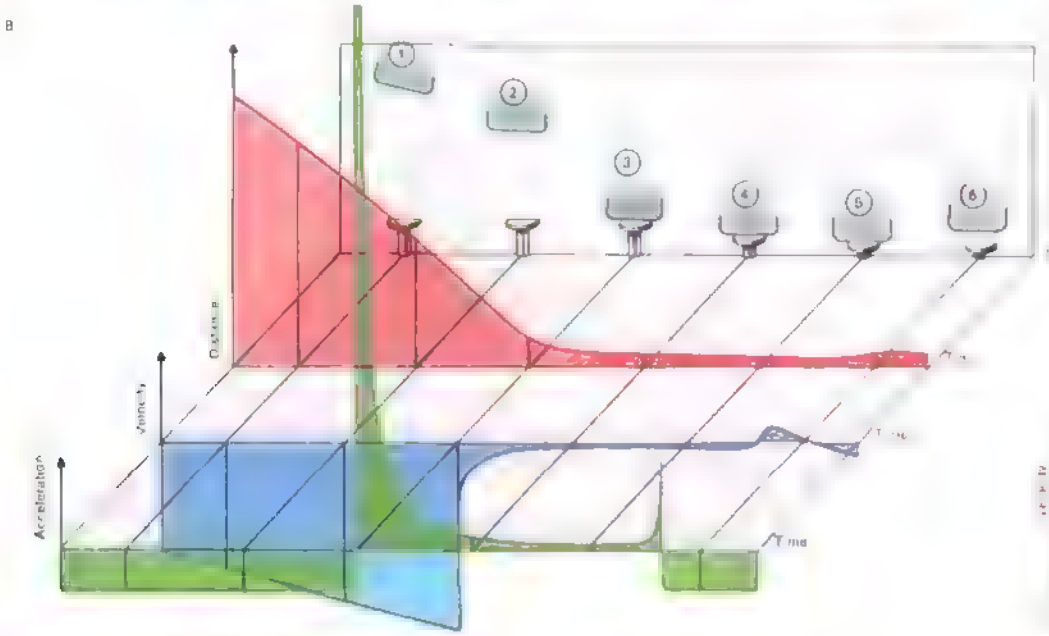
Electronic applications

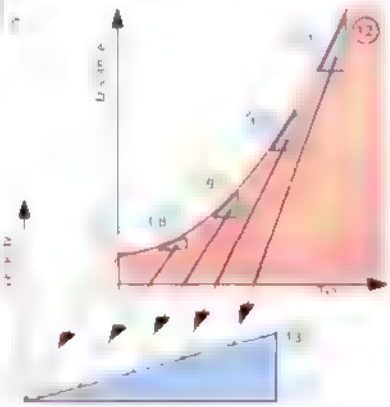
Analogous applications occur in electrical engineering [Key]. Take, for example, a resistor (across which the voltage is proportional to the current), a capacitor (in which

CONNECTIONS

See also







1 A falling hammer is shown [A] in six exposures 0 03sec apart and the hammer position curve (D in diagram B) follows them (1–6) Its changing slope is important and diagram C explains how this is measured like a road gradient, by tangent triangles with so many units vertically for each one horizontally The slopes at points 7–11 of line 12 are 0 2, 1 2, 2 2, 3 2 and 4 2 respectively – or 0, 1/2, 1, 1 1/2 and 2. Line 13 plots the increasing slope of line 12 (Obtaining a slope curve from a curve is called differentiation, the reverse is integration) In [B], the slope of the distance curve for the first exposure is given by a vertical drop of 2 8cm for a time interval, measured horizontally, of 0.02 sec So the velocity here is 2 8cm in 0 02sec = 140 cm/sec Similar slope measurements give the velocity at other points and yield velocity curve V which initially increases downwards, like the distance The acceleration (the rate of change of velocity with time) is the slope of the velocity curve – constant while the hammer is falling freely (curve A). At exposure 3 the hammer hits the nail at 200 cm/sec and things start to happen In a millisecond (thousandth of a second) the velocity is braked back to about 40 cm/sec (curve V), this rapid deceleration drives A off scale upwards to perhaps 100 times its free-fall value Newton's law then gives a correspondingly large force on the nail, 100 times the weight of the hammer, and this is how the hammer works – driven by the force generated by rapid deceleration of the head As the wood seizes the nail the shock kicks the hammer upwards, then in free fall again the hammer resumes steadily increasing downward velocity and constant acceleration

the current is proportional to the rate of change of voltage with time), and an inductor (where the voltage is proportional to the rate of change of current with time). Connect them all together and apply an alternating voltage. What happens? Calculus swiftly expresses this seemingly mind-boggling tangle as a differential equation and solves it to show, among other things, that at a certain frequency the whole affair "resonates" and very large currents can flow for very little applied voltage. Resonance is of fundamental value throughout electronics. The tuning control of a radio selects one station out of many by setting a circuit to resonate at the station's transmission frequency.

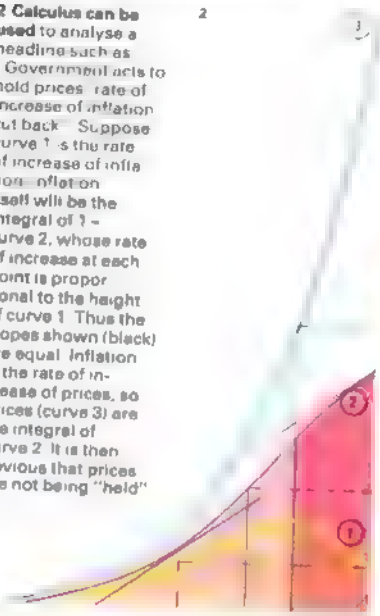
One powerful application of calculus is in seeking maxima, minima, and optima generally. A vertically thrown ball is momentarily stationary at the top of its flight: when the height is a maximum, the rate of change of height with time is zero. It can be found by differentiating the expression relating height and time and putting this equal to zero. This is a general rule of great value: for all technology is governed by the search for optima.

What is the optimum speed for a journey, for example? Too slow wastes time, too fast wastes fuel: both of these have an assignable cost. Calculus enables the rate of change of overall cost with speed to be found and the speed for which it is zero. This must then be the speed for minimum cost. Such calculations are essential in making the best use of ships and aircraft and the same principles apply in all searches for the best design or flow-rate or working temperature of almost every industrial system [6].

Universal principle

Many physical laws embody the same principle [3, 4, 5]. Thus light traverses an optical system by a path that takes less time than any other possible path — a principle from which the whole of classical optics can be derived. Indeed Leibniz, possibly carried away by the power of his creation, proposed that the whole universe had been designed in some such mighty self-optimization process and that this was the best of all possible worlds. He seems not to have considered the possibility that it might be the worst.

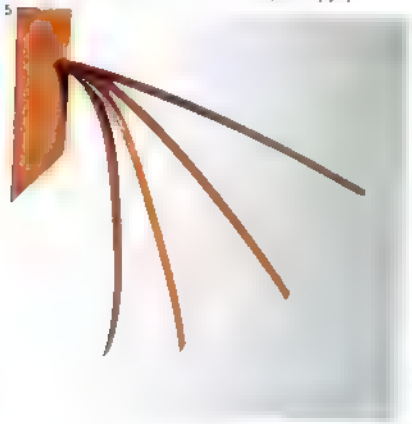
Faraday's "anchor ring" was the first ever transformer. It revealed the law that the voltage on the output coil depends on the rate of change of voltage on the input coil.



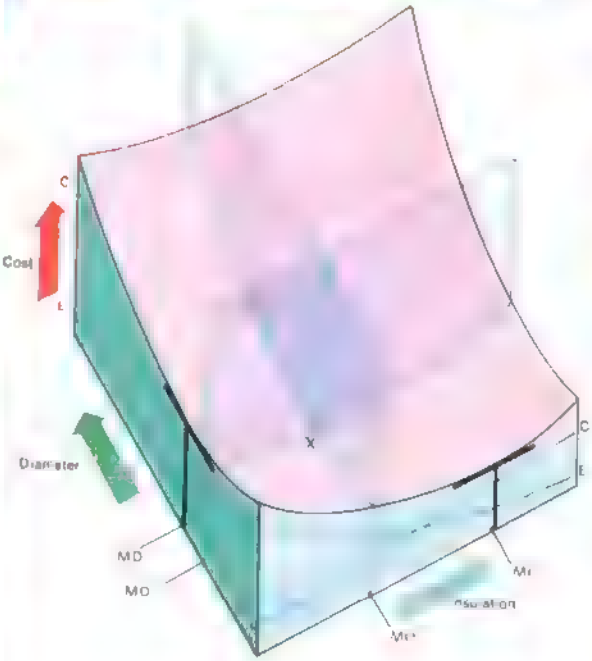
2 Calculus can be used to analyse a headline such as "Government acts to hold prices: rate of increase of inflation cut back". Suppose curve 1 is the rate of increase of inflation: inflation itself will be the integral of 1 – curve 2, whose rate of increase at each point is proportional to the height of curve 1. Thus the slopes shown (black) are equal. Inflation is the rate of increase of prices, so prices (curve 3) are the integral of curve 2. It is then obvious that prices are not being "held".

3 The stability of ships, buoys and other floating objects depends on whether a small tilt raises a greater weight than it allows to fall. In calculus terms stability requires a positive rate of change of height energy with tilt. The picture shows that with more than a critical volume of ballast liquid, a cylindrical tube floats vertically (1). Tubes 2 and 3 have too little.

4 A tower can be built by repeatedly jacking up what has been made so far and adding a lower most storey strong and heavy enough to support it all. Each new section must be bigger than the last to support all the previous sections and the rate at which successive sections get bigger will increase too. This type of growth (rate of growth increasing with state of growth) is called "exponential" and soon leads to explosive increase (A). The Eiffel Tower (B) is nearly exponential in form.



5 The curvature of a beam at any point depends on its loading. A projecting strip is bent at each point by the weight of the length beyond it and calculus-based beam-theory adds up all these changing curvatures to arrive at the final shape of self-loaded beams. These plastic strips have different thicknesses and their degrees of deflection vary roughly as the inverse square of their thickness. Engineering beams do not sag as much but the same design principles apply.



6 The cost benefit of increased insulation on a pipe (A) depends on reduction of heat loss (E on the pale blue) offsetting higher capital costs (C). Their sum is least at Mi. Increasing diameter reduces pumping costs (E on the dark blue) while raising capital cost (least at MD). The overall minimum cost is (Mi' MD') at X. A chemical plant (B) must minimize cost over hundreds of variables.

Lines and shapes: geometry

Imagine the United Nations decided to encircle the world at the equator with a steel band symbolizing international unity. If the contractor made it too long by one part in ten million – 4m in 40,075km (13ft in 24,900 miles), how high would it stand above the surface all round the globe? The answer is 63.7cm (approximately 25ins).

Lines and shapes working for a living

The foregoing is an example of simple geometry, the mathematics of size and shape. Since all solid objects have size and shape geometry is one of the most practical mathematical studies. If someone wants to know how thick to make a rotating shaft to transmit a certain amount of power, or what contour to give a ship's propeller, or even how much paint is needed to cover a room or concrete to lay a path he uses geometry to provide the solution [5]. Indeed, geometry arose from the surveying needs of the early Egyptians, who had to share out fairly the featureless hectares of fertile mud left by the annual flooding of the River Nile.

The Greeks took geometry over [3] and built an amazing intellectual edifice out of it. Euclid's *Elements of Geometry*, which was written in about 300 BC, develops a complete "axiomatic system" – a web of interlocking proofs all derived from a few basic axioms. "If you can't prove it, you don't know it!" challenged the *Elements* and ever since the admitted business of mathematicians has been the clarifying of basic axioms and the proving or disproving of statements derived from them.

A practical engineer seldom bothers with proofs, he generally accepts the mathematician's formulae and uses them. And almost instinctively, because geometry makes it simple, he designs objects from rigid parts linked at pivot joints. Many mechanisms around us embody the truths of plane geometry. The motions of a typewriter, the pantograph of an electric locomotive [2], the suspension of a car, the linkages in a sewing-machine or an autochange gramophone can all be described as working "models" of a set of geometrical theorems.

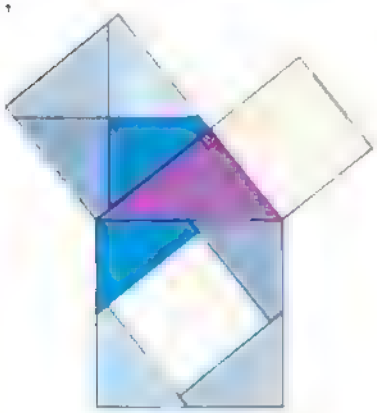
Some machines – printing presses and knitting machines for example – appear almost magical in the motions they generate by ingenious geometrical linkages. And most rigid structures use the geometrical fact that a triangle is the only rigid figure. A triangle of rods joined at pivots cannot deform whereas a square, say, can distort to a diamond shape. As a result, girderwork is generally made up of triangles (a big girder bridge is a good example, as is a geodesic dome [Key]).

Pi in the sky, and elsewhere

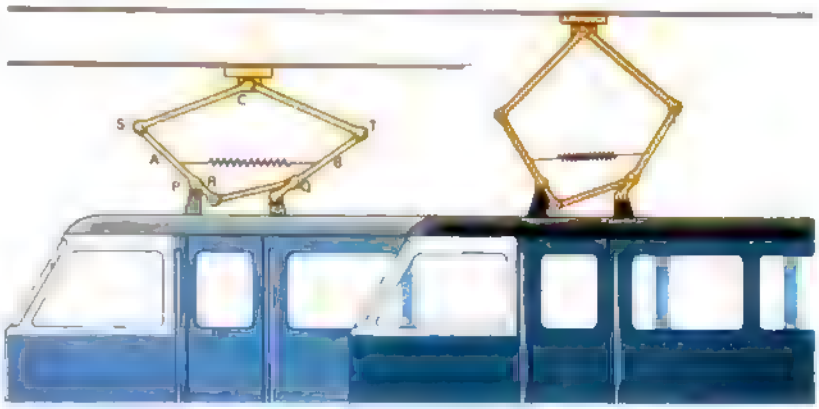
The circle is a simple geometrical shape but one that is mathematically rich. The Greeks succeeded in proving its circumference to be $2\pi r$ and its area πr^2, where r is its radius and π some number between $3\frac{1}{7}$ and $3\frac{10}{71}$. In fact π cannot be expressed as any whole number fraction. Expressed in decimals it begins 3.1415926535 . . . and goes on for ever, with a never-ending series of numbers after the decimal point with no numbers repeating. It is a fundamental constant in trigonometry, a numerical branch of geometry invented for mapping the stars and now fundamental to astronomy, navigation, surveying and all kinds of practical measure-

CONNECTIONS

See also

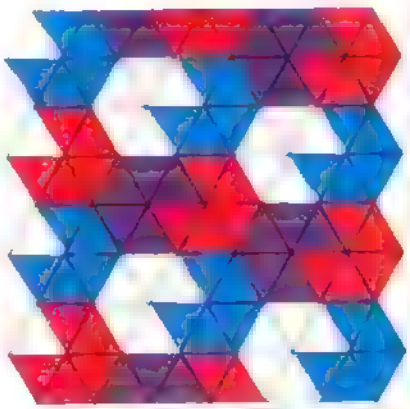


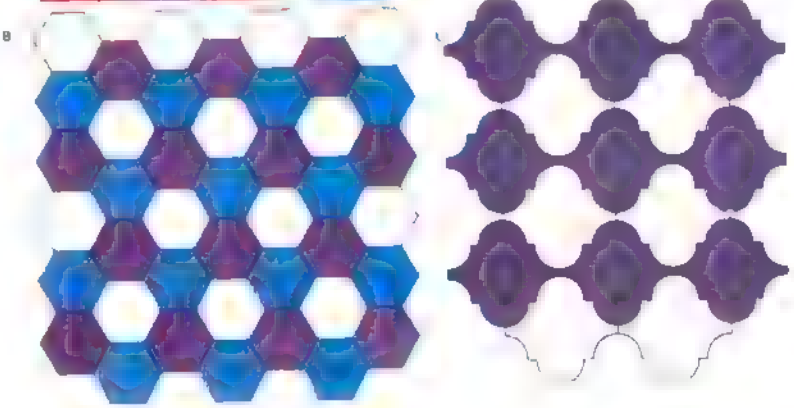
1 Pythagoras' theorem is the famous one children learn at school. The square on the hypotenuse (the longest side of a right angled triangle) is the sum of the squares on the other two sides. The big bottom square (the hypotenuse square) divides into four corner sections which can be re assembled into the top left square and the one in the centre, which is the size of the top right square.



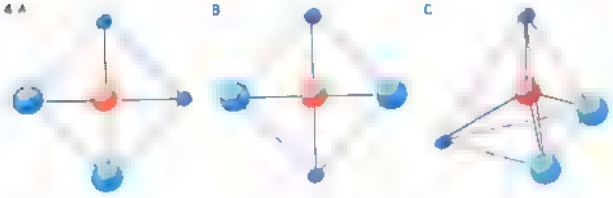
2 An electric locomotive must always maintain contact with the overhead wire which itself is not perfectly level. The geometry of the pantograph achieves this. A spring system urges points P and Q together, pivoting arms A and B to shorten distance ST. This distortion of triangle STC keeps C against the overhead wire. As arm A pivots linkage RQ makes B pivot too, preserving symmetry.

3 The ancient Greek Eratosthenes measured the earth's circumference by geometry. He found that when the sun was overhead at Syene it was 7° from the vertical at Alexandria. He knew the distance between them, about 800km (500 miles), and he reasoned that it represented 7° at the earth's centre. The full 360° of circle representing the earth's circumference must be 360 ÷ 7 × 800 (500) = about 41,140km (25,700 miles).





5 Tiling a floor with identical tiles can be done in various ways. Obviously it can be done with equilateral triangles (A), hexagons (B), squares, or with tiles made by fusing shapes together (C). But it cannot be done with pentagons or any tiles with pentagonal symmetry. Geometry proves that there are just 5 basically different tiling patterns – including the most ornate using identical tiles.



4 Molecules are too small to be visible so chemists use geometry to deduce molecular structures. Dichloromethane (a solvent used in paint-strippers) has one carbon atom, two hydrogen atoms and two chlorine atoms in its molecule. If they were arranged as a square with the carbon in the middle, two forms of dichloromethane should exist, one with the chlorines adjacent (A) and one with them opposite (B). But if the atoms are arranged tetrahedrally only one form is possible (C). Only one form has ever been obtained so the square structure is wrong. By such reasoning chemists deduced the spatial arrangements of thousands of molecules long before methods such as crystallography provided direct evidence.

ment. In fact π has 'escaped' from geometry and pervades all numerical measurements.

Some of the most elaborate geometry based on the circle is used in lens design. Almost all lenses – for cameras, spectacles, telescopes and so on – have circular cross-sections. Tracing the light path through a multi-component lens system is a complex geometrical task now carried out by computers. The computer programs calculate the characteristics of many possible lens designs and select the one with the fewest aberrations (for no lens system can be absolutely perfect). The result is a compromise, but the best that can be reached bearing in mind the practical difficulty of actually grinding the lenses.

Geometries beyond intuition

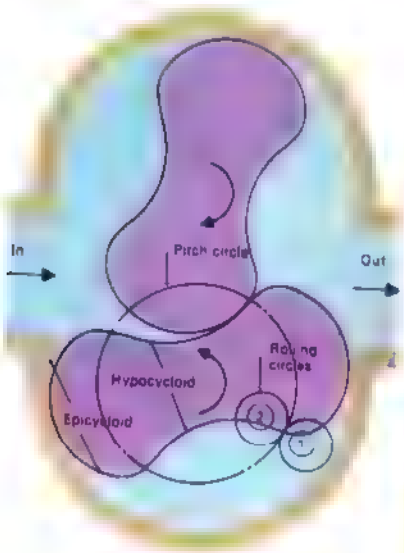
Euclidean geometry takes a number of intuitive notions for granted – the idea of a straight line, for example. Euclid thought of it as a line of zero curvature, the shortest line that could be drawn between two points. In practical matters, such as sighting and surveying, we assume that light travels in straight lines. But the physicist feels free to question these suppositions. He considers it possible that light flashed out from the earth might go all round the universe and return to its starting-point, just as a person would who travelled in what he regarded as a "straight line" on the spherical earth. Indeed cosmology, the study of the universe as a whole, currently favours a "closed curved" universe with a finite volume but no boundaries just as the earth's surface has a finite area but is without edges.

Mathematicians see Euclidean geometry as just one of many imaginable geometries, each true of space of a particular curvature [10]. Their theorems may be strange, but provided they can be rigorously derived from the stated axioms (assumed facts) mathematical protocol is satisfied. And which of them is true of our real space is a matter of scientific experiment, not of axiomatic assertion. Fortunately, any curvature must be very small, so that Euclidean geometry works well in the small volumes we can deal with, just as in mapping a small area of the earth it can be assumed to be flat without significant error.

KEY

A geodesic dome is a rigid structure made from many tri angles and designed for both lightness and strength

6

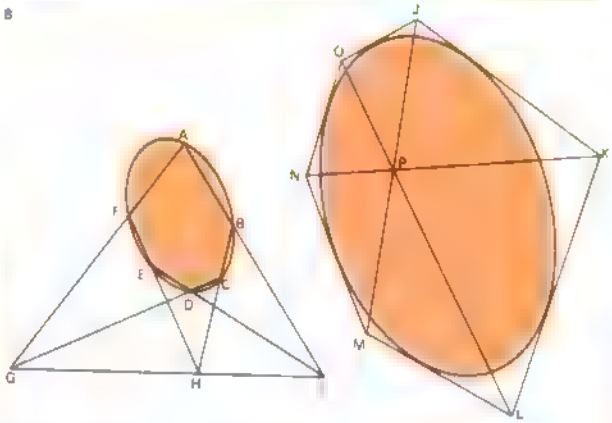


6 An air compressor uses the subtle geometry of interlinked cycloids. The end-lobes of each "paddle" have the curve traced by a point on a small circle [1] rolling outside the pitch circle. Its waist has the curves from a similar circle [2] rolling on the inside of the pitch circle. As the paddles mesh like 2-toothed cogwheels, they always touch each other trapping successive volumes of air and compressing them.

7

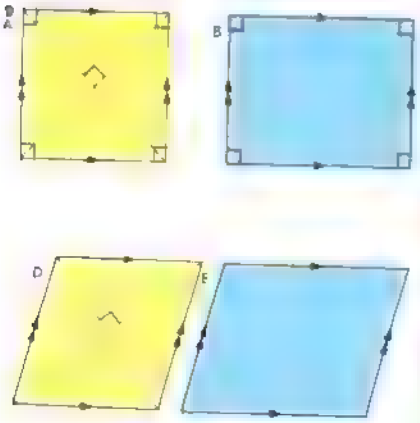
7 This model of Felton and Murray's early steam engine is a geometrical theorem in action. The inner cogwheel rolls around a fixed outer gear of twice its diameter. Geometrically, this implies that one point on it reciprocates in an exact straight line. A piston rod attached at this point is driven from a cylinder, a crank takes the drive from the centre of the rolling wheel, to drive other machines.

8



8 The principle of duality in geometry states that any two lines define a point (their intersection) and any two points define a line (the one joining them). If six points [A–F] touch an ellipse, then the lines joining them form three opposite pairs whose points of intersection meet at a single line [GHI]. The dual of the theorem is that if six lines touch an ellipse then the points at which they intersect [J–O] form three opposite pairs whose lines of connection meet at a single point [P]. So points and lines are "duals" of each other and if these words are interchanged in a theorem, a new theorem results.

9



9 Quadrilaterals include a square [A], with right-angles and all its sides equal and parallel, a rectangle [B] with only opposite sides equal, a trapezium [C] with only two opposite sides parallel, a rhombus [D] and a parallelogram [E], both with opposite sides parallel and no right angles.

10

10 Euclidean geometry [A] is not inevitable and may not be true of real space. Mathematicians accept any geometry that is not self-contradictory and recognize many different kinds. In Lobachevskian geometry [B] the angles never reach 180°, like geometry on a trumpet surface. In Riemann geometry [C] angles of a triangle always exceed the Euclidean 180°, like geometry on the surface of a sphere. Extended to three dimensions this is geometry of "curved space".

Lines and angles: trigonometry

The Simplon tunnel between Italy and Switzerland, is 20km (about 12 miles) long and was bored from both ends through the Alps. When the headings met in the middle in 1906, they were in exact horizontal alignment and only 10cm (4in) out vertically. The engineers managed to smooth out the discontinuity. Using trigonometry they had set up their machines to cut along the 10km (6 mile) sides of two huge triangles in the mountain.

Sines, cosines and tangents

Trigonometry is the art of calculating the dimensions of triangles. The basic idea [1] is that the ratios between the sides of a right-angled triangle depend on its base angle [A]. The ratios have been named the sine of A (sin A), the cosine of A (cos A), the tangent of A (tan A) and others. They have been tabulated for many values of the angle A. Sin A is the length of the triangle side opposite the angle A divided by the longest side; cos A is the length of the side adjacent to the angle A divided by the longest side, and tan A is the ratio of the length of the opposite and adjacent sides of the triangle.

Armed with trigonometrical tables anyone can determine the dimensions of any triangle with great accuracy. Since nearly any shape can be broken up into a series of triangles this is a powerful method of solving even complex spatial problems. To use it in tunnelling engineers set up a station from which both the ends are visible or (as this may be difficult with mountains all around) a station from which other stations are visible from which in turn the ends can be seen. They measure the angles between all the stations by optical sighting and thus relate the two ends. Trigonometry then tells them the tunnelling angles that will align the two headings. The required accuracy of a thousandth of a degree implies a certain expertise; but the mathematical principle involved is nevertheless extremely simple.

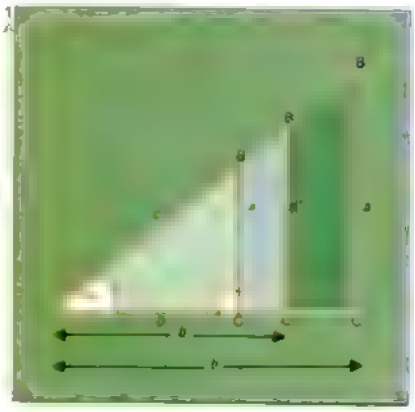
Trigonometry in everyday life

Trigonometrical ratios have, however, 'escaped' from their simple geometrical interpretation and uses in surveying and measuring, and now crop up in all sorts of mathematical problems that do not seem to be at all "angular". Some of their most fruitful applications are in circuit theory, radiation physics and information-handling, in which the angles are not real but introduced merely for convenience.

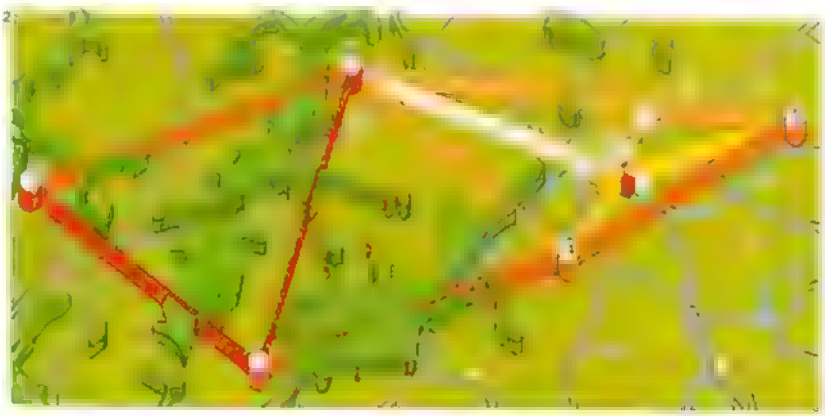
The sine of 0° is 0 and it increases with increasing angles up to 90°, whose sine is 1. Between 90° and 180° the sine reduces again to 0. From 180° to 270° the sine is negative decreasing to −1. And from 270° to 360° the sine increases again from −1 to 0. Thus if a trigonometrical angle is regarded as winding up continuously [5], its sine swings between +1 and −1 and back at each revolution of 360°. This periodic behaviour gives mathematicians a framework for handling waves, vibrations, oscillating radiation such as light and radio waves, and alternating current (AC) electricity. In most European countries a power station generator spins at 50 revolutions a second. As a result its output voltage (which depends on the sine of the angle of rotation) swings back and forth between positive and negative at 50 cycles per second (50Hz) to generate mains-frequency AC. Any other source of oscillation, even

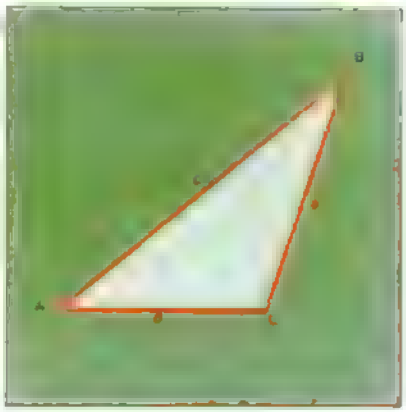
CONNECTIONS

See also



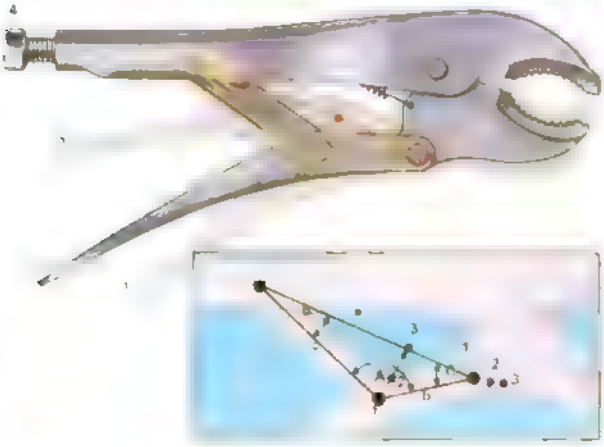
1 Trigonometry (literally, "triangle-measuring") is based on six named ratios in a right-angled triangle. Triangles ABC, AB'C', AB''C'' all have base angle A. Clearly they have the same relative proportions so $a/b = a'/b' = a''/b''$. In fact any right-angled triangle of base angle A will have this fixed ratio between those sides. It is called the tangent of A or tan A. Thus when A = 45°, $a = b$, so tan 45° = 1. The other ratios are b/a, cotangent (cot A); a/c, sine (sin A); b/c, cosine (cos A); c/b, secant (sec A); c/a, cosecant (cosec A). These are tabulated for all angles nowadays some pocket calculators can work them out. They will give the dimensions of any triangle not just right-angled ones, via the formulae (B): $a/\sin A = b/\sin B = c/\sin C$ and $a^2 = b^2 + c^2 - 2bc \cos A$. (Side a is always opposite angle A, side b opposite angle B, etc.)



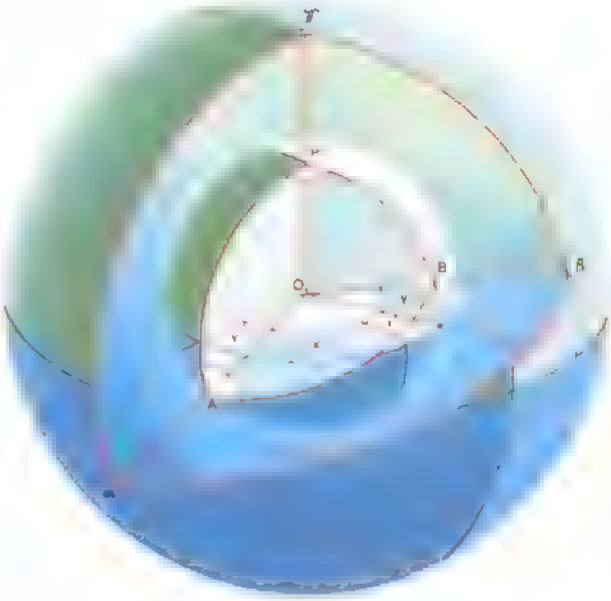


2 Surveying by "triangulation" uses the formulae that fix any triangle if one side and two angles are known. Distance (1–2) is carefully measured as the fundamental base line. Reference point (3) is selected and the angles of triangle 123 determined by optical sighting. This fixes point 3 and enables distance 2–3 to be calculated. Sighting from these established reference points will then locate any others (4, 5,

3 On a sphere such as the earth any distance can be represented by the angle it makes at the centre. Thus distance PB may be represented by angle POB. Accordingly positions are defined by angles of latitude (with the equator) and of longitude (north/south line). Point A has longitude x° west, latitude y° south, B has x'' east, y'' north. The "spherical trigonometry" of "spherical triangles" such as PAB tells a navigator the distance AB and the compass bearing (angle A) of the journey. Similarly in mapping the heavens astronomers locate stars on spherical celestial triangles like $\pi\alpha\beta$.



4 In triangle ABC of the self-grip wrench $a^2 = b^2 + c^2 - 2bc \cos A$. As b and c are constant lengths a^2 and hence a, change only with the cos A term. Reducing A increases cos A and hence a, closing the jaws. When A is large, a small decrease in A substantially. But as A approaches 0° cos A is near its limit of 1 and changes little. The fine closure moves the jaws slightly giving a big leverage ratio with gripping force.



light with a frequency of 600 million million Hz, can be similarly assigned a notional "phase angle" winding up at the appropriate rate of time

Any vibration, however complicated, can be made up of a set of sine-wave components (or cosine-wave ones which are similar), each with its own frequency. Each frequency is quite independent of the rest. (Two stones thrown into water together generate two sets of spreading ripples which intersect and go right through each other, emerging quite unaffected.) Similarly, the human ear can pick out the notes in a chord although they make a single vibrational pattern in the air or a single groove on a gramophone record

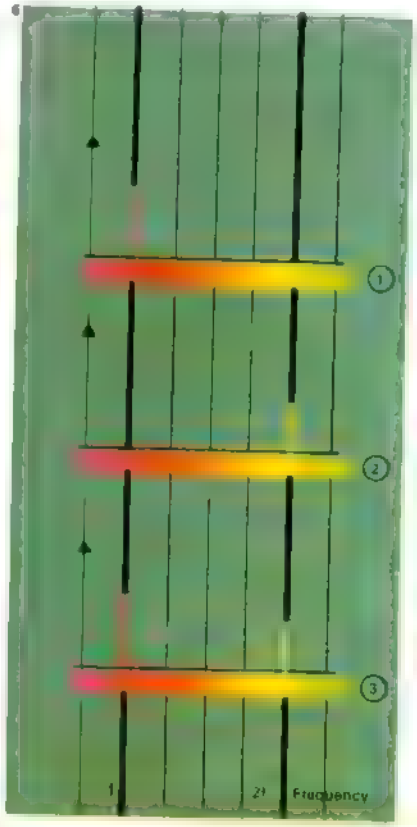
Angles in a radio beam

Many electronic techniques process these frequency components of vibrations in ways governed by trigonometry. An AM (amplitude modulated) radio transmitter, for example, has to take a sine-wave audio frequency A (say the musical note A, 440Hz) and attach it somehow to a radio sine-wave "carrier" C, being broadcast at perhaps one million Hz (1MHz, in the medium-wave band). It does this in effect by multiplying the audio voltage at each instant by the carrier voltage at that instant and transmitting the result. Now one of the many trigonometrical formulae for simple angles asserts that $\sin A \times \sin C = \frac{1}{2}\cos(A-C) - \frac{1}{2}\cos(A+C)$. Since A and C are phase-angles of audio and carrier frequencies the result of the multiplication is two cosine-waves (just like sine-waves), one at (1,000,000 − 440) Hz and the other at (1,000,000 + 440) Hz, each of half the intensity of the original carrier

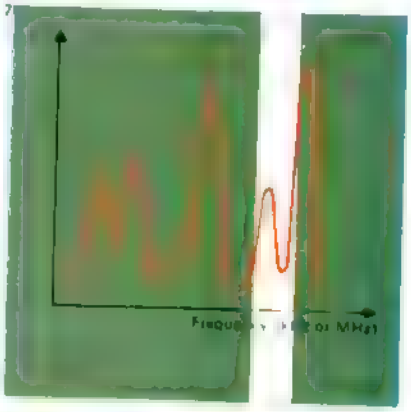
The splitting of the carrier into two closely spaced "sidebands" is called amplitude modulation, or AM. A transmission generally has many such pairs of sidebands continuously changing in their spacing and intensity with the changing frequency-components of the audio signal. At the receiver the audio signal is recovered by the reverse process of demodulation. It may seem incredible that a mathematical formula first proved for static triangles on paper can be impudently applied to the imaginary rotating angles of an electronic signal

The quadrant was an early instrument used by astronomers to find the altitude of the heavenly bodies. The surveyor's quadrant developed as a portable version for surveying and artillery ranging. This example was made by Jacob Lusuerg of Rome in 1674. Its most interesting feature is the Vernier scale – invented by Pierre Vernier (c. 1580–1637) in 1631 – for measuring to 1/60°. This is the lower arc-scale joining the legs of the pivoting V-shaped unit which slides over the static quadrant base plate. Another scale shows the tangent of the measured angle

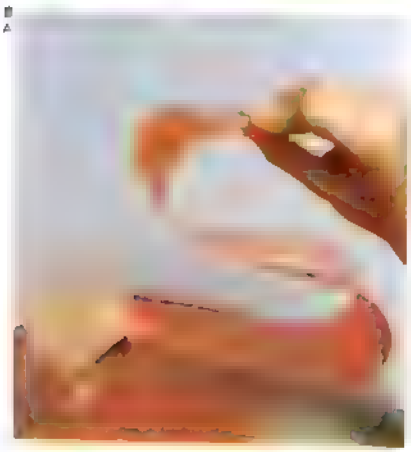
5 As a rotating radius sweeps out an ever increasing angle, the angle's sine varies cyclically, repeating itself for every additional 360° of rotation. For a circle of unit radius the sine is the height of the end of the radius above the horizontal. Such sinusoidal waveforms occur in vibrations. The frequency is the number of radius rotations per second. Two simultaneous sine-waves of different frequency will add together to a complex waveform thus sine-waves [1] and [2] add to give the waveform [3], which might represent the variation in sound-pressure of two notes sounding together



6 A waveform can be made by combining sine-waves. It can also be broken down into them. This diagram shows the amplitude (intensity) spectra of the waveforms of illustration 5. Waveform 1 has only one component in its spectrum, at the frequency f. Waveform 2 has a single component of frequency 2f but of lower intensity. Their combination [3] has both these lines in its spectrum



7 The complex waveform of all the signals entering the radio receiver's aerial will have many components in its frequency spectrum. Each peak is a broadcast on a specific frequency. Some stations are weak, some strong, tuning the radio moves a narrow frequency-acceptance band along the frequency scale to select just one of them. The small modulation is then decoded to give sound



8 A bent steel strip [A] does not adopt a sine-wave form but a related "sine generated" curve. The direction of the strip from point to point varies sinusoidally with distance along the strip. This minimizes the energy of bending stored in the steel. This curve is created on a grand scale [B] when a slow-moving river winds to the sea. The water has little energy and so seeks the line of least resistance as it cuts its channel.

Surface and volumes: solid geometry

In 1826 the German astronomer Heinrich Olbers (1758–1840) asked what may seem to be a silly question: Why is it dark at night? Silly questions are sometimes the most profound and Olbers tackled this one using straightforward solid geometry. He imagined the universe divided into a series of concentric shells around the Earth, like the layers of an onion spreading out to infinity. He supposed that the stars were more or less uniformly distributed. Then, through solid geometry, Olbers calculated that a shell twice as far away is bigger and contains four times as many stars. But, in theory, only a quarter as much of their light should reach the Earth. Each shell therefore contributes the same radiance to the night sky no matter how far away it is. Because there are an infinite number of shells the night sky should be infinitely bright – or at least as bright as the face of the Sun.

So why is the night dark? Even today, astronomers do not agree on the structure of the heavens. The universe might be finite in space, with only a limited number of shells, or finite in time, so that light from the most distant shells has yet to reach here; or it may be expanding, weakening the light from distant shells. Olbers's paradox remains an outstanding example of how simple mathematics can provoke the most surprising conclusions from uncontroversial assumptions.

Sizing things up

Mathematicians and engineers alike have to be able to calculate the areas and volumes of various solid objects. For an object with flat faces the surface area equals the sum of the areas of the faces. Thus for a cube, the surface area is merely six times the area of one face. For a sphere the area is four times π times the square of the radius. The volume of a cube is the length of one side multiplied by itself three times (the length cubed), and the volume of a sphere is 4/3 times π times the cube of the radius.

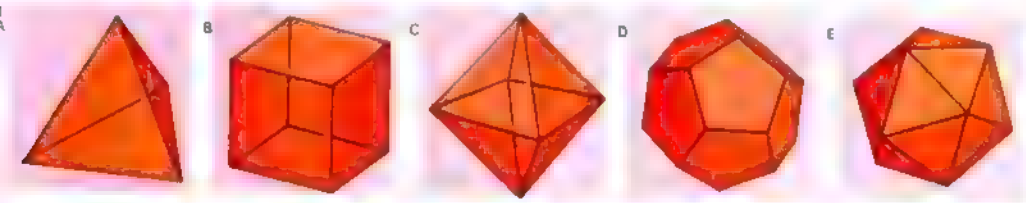
Pyramids, prisms, cylinders, cones and ellipsoids present more complex problems, but all can be calculated using solid geometry, that is the geometry of shapes in three dimensions. Mathematicians use solid geometry to find the surface areas of such shapes and whether they can be made by forming a flat paper shape (a cylinder, prism, pyramid and cone can be made this way, but not a sphere or ellipsoid). The path that a grinding-wheel of known dimensions must traverse to cut a given shape from a metal blank, or how much earth must be shifted to make a railway embankment of a given height, or what size cylinders can be bored in an engine-block for a given safe spacing between them (and the resulting swept volume) – to determine all these quantities, engineers make use of solid geometry.

Networks of force

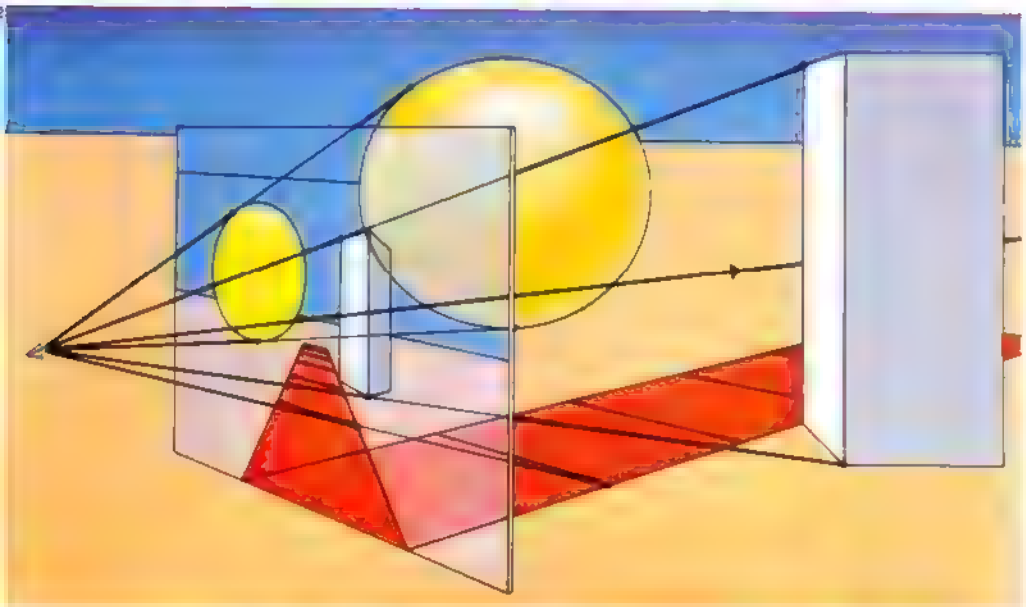
The subject-matter of solid geometry includes not just the shapes of objects and assemblies, but the invisible strains and forces that traverse them. The centre of gravity of a cylinder is half-way up it; stood on end and tilted it will not fall over provided that any part of the top surface is still vertically above any part of the bottom surface. But the centre of gravity of a cone is a quarter of the way up it. It can be tilted until its tip is one-and-a-half times as far to one side as the

CONNECTIONS

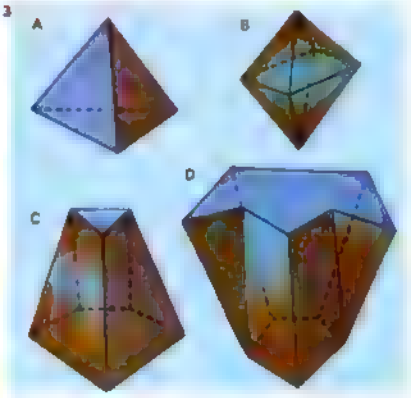
See also



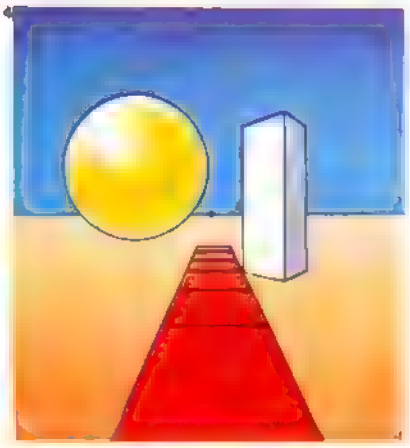
1 A regular polygon has all its sides and angles equal, as in the equilateral triangle, square and pentagon. Euclid proved that there can be only five regular solids whose faces are all identical regular polygons: the tetrahedron (A), the cube (B), the octahedron (C), the dodecahedron (D) (with 12 faces), and the icosahedron (E) (with 20 faces). These can be found in nature as crystals and viruses.



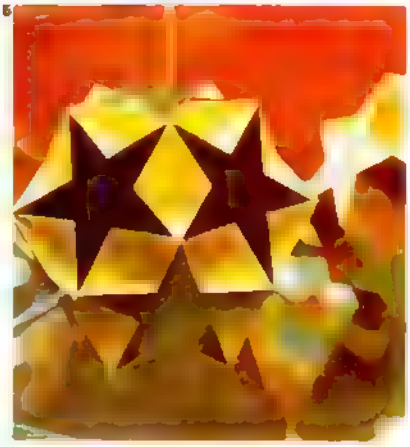
2 Solid geometry controls the perspective appearance of the world because light travels in straight lines. The laws of perspective envisage a picture-plane between the eye and the scene to be represented. Connect each point in the scene to the eye by a straight line, the place where this penetrates the picture-plane is its position in the perspective representation of the scene. From the eye's viewpoint, a perspective picture is seen



3 All solids that do not have holes through them, and have flat faces, obey Euler's theorem $V + F = E + 2$ where V is the number of vertices (corners), F the number of faces and E the number of edges. For the tetrahedron (A) $4 + 4 = 6 + 2$; for the octahedron (B), $6 + 8 = 12 + 2$. The shapes C, D also obey the rules. In fact the shape and size of the solid does not matter at all.



4 A perspective picture has a central "vanishing point" to which parallel lines perpendicular to the picture-plane all converge. Other horizontal lines (such as edges of the cuboid box) converge to other points on the picture's horizon. A circular disc is distorted by perspective representation unless it is directly in front of the eye: the effect is small for most deviations from this position. This picture is the perspective view of illustration 2



5 Uniform polyhedra can have several different regular polygons contributing to their faces. There are 13 "Archimedean solids", not counting the infinity of simple prisms allowed by this definition – each of which has a regular polygon top and bottom joined by square faces around the middle. If faces are allowed to intersect, 53 additional uniform polyhedra result. This ditrigonal dodecahedron is composed of several dodecagons

edge of the base extends on the other side

Such simple results, elaborated for far more complex shapes, determine for example what form a dam must have in order that the water pressure should not push it over; how high in the water a boat of given shape will float, and how far over it will heel if loaded lop-sided; and what overloading of a tower-crane will just topple it

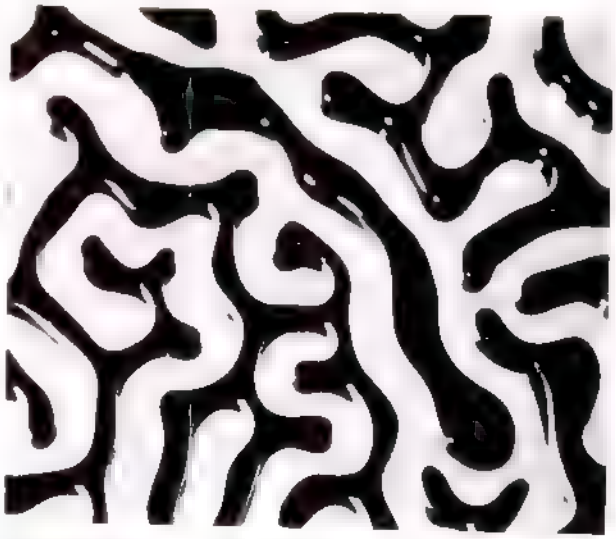
For forces more complex than gravity still more intricate questions arise What pressure can a round-ended gas cylinder withstand, and where will it fail if overpressurized? (Answer: on the inner surface, at a point midway between the ends.) What structure must an aircraft wing be given in order that, when loaded by lift and thrust and weight and drag, it will deflect into the desired shape without overstressing any of its parts? Problems such as these can be solved by modelling, or by computing and translating the solid geometry of the model into its numerical equivalents

There are systems whose geometry reveals the active forces directly. A magnetic-liquid labyrinth [Key] reveals the opposing magnetic forces that mould it. And in nature, a bone or a tree that grows against the forces on it reveals those forces to the intelligent eye by the shape it grows into – the ideal shape for the loads it has to bear.

Molecular architecture

The solid geometry of molecules is surprisingly important in modern chemistry It determines not only how they pack into crystals, but how they react. It is particularly significant for understanding enzymes, the powerful biological catalysts that bring about reactions which the chemist is often helpless to imitate. An enzyme is a huge molecule with a complicated active surface on which only the right reacting molecules can fit. And having fitted they are then held in the right positions to react. In doing so the reactants' geometry alters and they spring from the surface, leaving it ready to accept more reagents The double helix of the DNA molecule consists of two interlinked twisting strands. The whole marvellous mechanism of the human body depends on the sub-microscopic solid geometry of the fundamental catalysts of life.

KEY

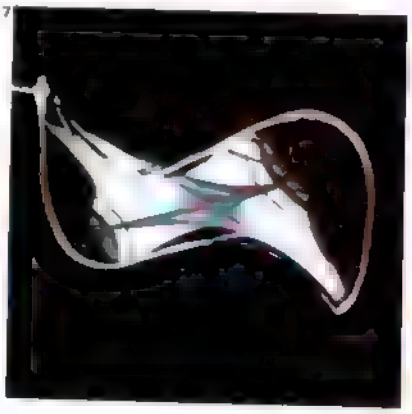


This mixture of a magnetizable liquid and an immiscible transparent one is in a magnetic field Every part of the magnetic liquid then repels every other part, so it seeks to divide into many small sections. But every division uses energy so the liquid compromises to the shape shown here.

6

6 The geometrical shapes of engineering objects are often beautiful This sludge-pump impeller is a stack of discs whose centres are helically disposed around the central axis of the impeller

7



7 There is no complete mathematical construction to the general problem (given a closed line) of finding a minimum-area surface that has that line as a boundary. A soap film solves it automatically for any closed line. The film is always in tension and shapes itself to minimize its area Here is one outlining the smooth and elegant minimum-area surface for a three-lobed loop made of copper wire

8

8 Optical components have surfaces governed by the laws of optics and the solid geometry of ray paths. The big rectangular outline mirror of this infra-red spectrometer has an ellipsoidal surface. It intercepts a diverging beam of infra-red radiation coming from the left and reflects it so that it all converges on the radiation detector, that is the stalk mounted in front of the mirror The mirror (bottom left) has a spherical surface and processes the invisible beam at an earlier stage of its journey through the instrument. The complete instrument has about 20 reflecting surfaces

C

E

D

9 A

B

9 For maximum volume from a given area of tin-plate, the height should equal the diameter A standard 425g (15oz) can [D] has a height 1 4 times the diameter because stamping out and forming the ends wastes some tin plate. This is thus the practical optimum The polish tins [A, B] are wider than ideal for easy access to the contents; the aerosols [C, E] are narrower to resist internal pressure

10

10 The absolute minimum surface area for a given volume container is a sphere (A soap bubble proves this by minimizing its surface area round an enclosed volume of gas) This also gives the greatest resistance to internal pressure Spherical tanks are used to store liquids under pressure Such tanks are also used for liquids held at low temperatures where the absolute minimum wall area minimizes leakage of heat from outside into the liquid

Shape and symmetry

If in the same boarding houses there are two rooms of equal aspect and furnishings, then their rents will be equal. For suppose the rents are different then one tenant will be paying less than he might, which is absurd This elegant theorem formulated by the Canadian humorist Stephen Leacock (1869–1944) illustrates the mathematical idea of symmetry

Symmetry is a powerful concept and its workings can be seen in many aspects of the world. The two halves of a bridge span, the wings of a bird or of an aircraft, the blades of a propeller, all have symmetry – for otherwise one of them is worse than it need be, which is also absurd. Mathematicians recognize many different types of symmetry all described by the group of real or imaginable "symmetry operations" which leave the symmetrical entity apparently unchanged A square, a cube or a four-bladed propeller can all be turned through 90 degrees without apparent change. they are said to have a "fourfold axis of symmetry". An irregular object has the lowest symmetry because any twist or turn is detectable A sphere has the highest possible symmetry, no twist or turn is detectable. This made it the "perfect" figure to the ancient Greeks and makes it highly useful today A ballrace is so simple because the balls need no aligning, no matter how they roll they cannot jam the bearing. A roller-bearing of lower symmetry needs guides to keep rollers parallel to the bearing axis; a tapered roller bearing of lower symmetry has even more geometrical constraints

Symmetry of nature

A snowflake [Key] shows how the laws of nature give symmetry to their products. It has 120-degree angles between many faces because in the water molecules of which it is comprised two hydrogen atoms form a 120-degree angle with an oxygen atom. The crystal lattice in ice is formed by the regular interpacking of the molecules and reflects this symmetry

But this does not explain why the whole elaborate structure has a sixfold axis of symmetry How does one branch of the flake know how its fellows are growing, so as to imitate them exactly? The physicist Samuel Tolansky (1907–73) made the suggestion that a snowflake, as it falls and takes up water vapour from the cold air, is vibrating with the symmetry of its crystal structure. All the branches move and twist together in a complex and changing pattern; the fastest points on each branch intercept the most water vapour and so grow together

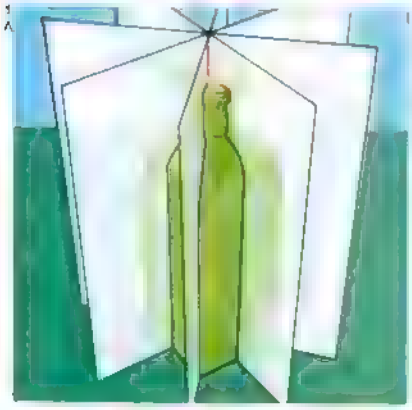
Such symmetries of process are common throughout nature. The radial shatter pattern of a broken window betrays the symmetrical stresses that radiated outwards from the impact point

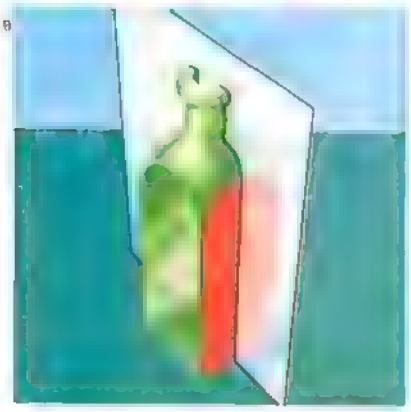
Symmetry in the abstract

Mathematics manages symmetry by "group theory", a fascinating topic which, from a few apparently trivial axioms, develops rapidly into a structure of amazing subtlety and elegance. The oddest thing about it is that unlike number theory, it allows $a \times b$ not to equal $b \times a$. This lack of symmetry in the mathematics of symmetry may seem like complete nonsense, but in practice the order of events can also be important. Sanding down a door, then painting it, for example,

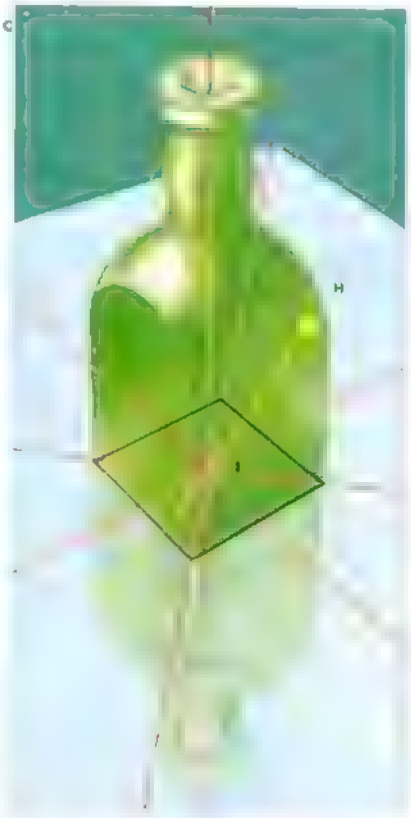
CONNECTIONS

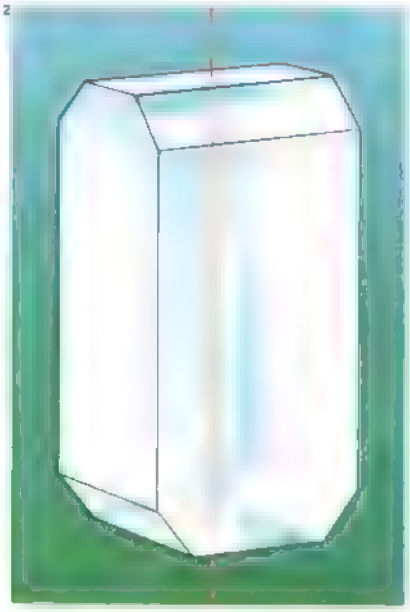
See also



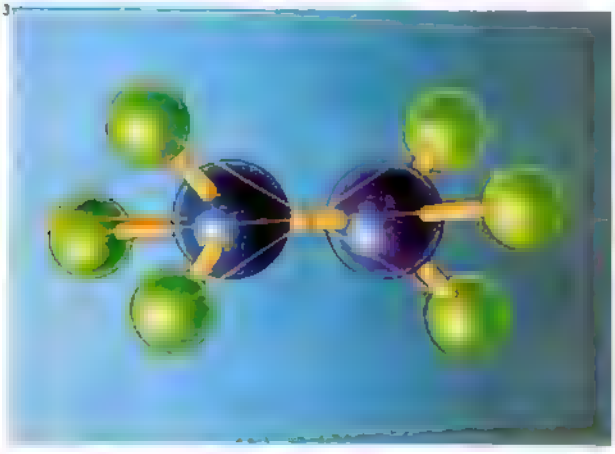


1 An object or mathematical entity has symmetry if some defined "symmetry operation" on it leaves it unchanged If the bottle [A] is rotated through 90° about its vertical axis it presents its original appearance again Because this symmetry operation would recur four times in a complete revolution the axis is a "fourfold axis of rotational symmetry" The bottle's other "symmetry elements" are four mirror planes Reflecting every point on the bottle through such a plane to the corresponding position on the other side is a symmetry operation. A bottle with a blank label [B] has no rotation axis and only one mirror plane as symmetry element The pseudo bottle [C] has new symmetry elements a horizontal mirror-plane H and four twofold axes in it, as well as a centre of inversion [I], about which the bottle can be rotated and remain unchanged





2 Crystals of chemical substances show symmetry that derives from the lattice of molecules composing them. The urea crystal has a vertical "improper axis of symmetry" for a symmetry element This means that rotating the crystal through 90° and then reflecting it in a bisecting horizontal plane leaves it apparently unchanged It also has two mirror planes and two 2-fold axes.



3 Ethane is a gas whose molecules each have two carbon atoms [blue spheres] and six hydrogen atoms [green spheres] It has the centre of inversion [I] but the two CH_3 groups are not related by a mirror plane between them. These groups can mutually rotate into alignment, giving the molecule such a mirror-plane but destroying the centre of inversion Molecular symmetry and its alteration with internal motion dominates much of chemical theory

4 The common starfish (*Asterias rubens*) has five appropriate planes of symmetry and a 5-fold rotation axis. Among animals only a few specialized sea creatures (radiata) have such high symmetry They probably evolved from ancestors of lower symmetry, as inferred from their larvae, which have the approximate mirror-symmetry of most creatures, including man. The starfish has no horizontal mirror plane, it has a true "top" and a true "bottom

gives a different result from painting it and then sanding it down

For an object with symmetry its group consists of the "operations" that can be carried out on it, turning it through 90 degrees, reflecting it in a plane and so on. Take a squat square-shouldered pill bottle without its label, hold it upright and pivot it through a right-angle about its top-left and bottom right corners. It will then be horizontal with the neck on the left. Turn it clockwise through 90 degrees and it will be upright again. That is a symmetry operation. But if the latter is performed first and then the former the bottle will finish upside down. $a \times b$ does not equal $b \times a$

The uses of group theory

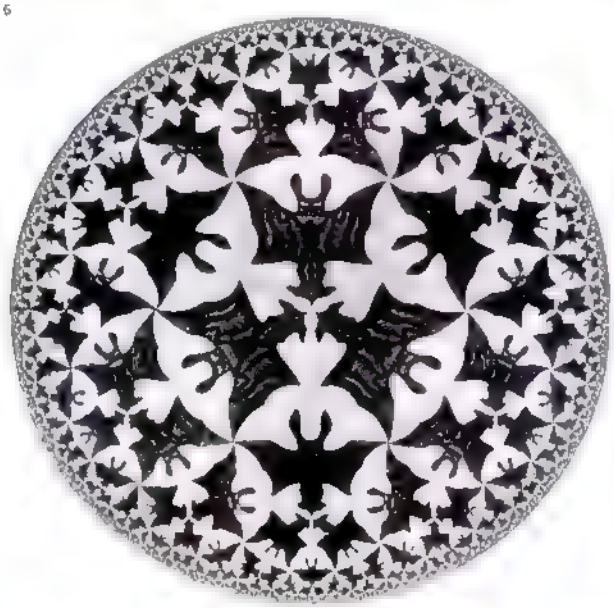
Group theory is one of the many inventions of nineteenth-century mathematics that later found scientific use. Indeed, the rapid spread of its strange but potent "arithmetic" in twentieth-century physics earned it the title, among an older generation of physicists, of *die Gruppenpest* ("group nuisance"). But its incorporation into modern physics and chemistry, with their need to understand the subtle symmetries of molecules and crystals and their energy states, has made possible the theories which give us such modern marvels as semiconductor electronics

So much symmetry exists in scientific theories and mathematics that researchers acquire a feeling for it and any "lopsided" features of a theory or experiment make them uneasy. In electromagnetism, for example, the fact that electric charges (positive and negative) can be isolated, whereas magnetic poles (north and south) cannot, seems somehow to be "wrong". Many physicists have sought magnetic monopoles to complete the symmetry of the situation but, so far, attempts to discover such particles have been unsuccessful. But the most daring of all such insights was that of Albert Einstein (1879–1955) when he reasoned that the speed of light (and indeed every phenomenon of physics) must be the same for all observers, no matter how fast they themselves were travelling. Implicit in that mighty assertion of symmetry was nuclear power and the atomic bomb

KEY

The symmetry of a snowflake echoes the symmetry of its molecules but owes its elaborate perfection to the subtle process of crystal growth by vapour-deposition on a vibrating surface

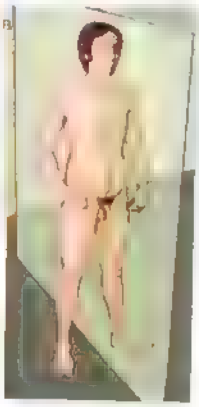
6



5 Just as an object can have symmetry, so can an infinite repeating lattice. The symmetry operations for objects also apply to lattices. But there are other operations which, applied to an infinite lattice of appropriate symmetry will leave it apparently unchanged. One is "translation" that is, shifting the lattice sideways. Every lattice can be divided into repeating "unit-cells" and displacement by one unit-cell spacing is a symmetry operation. Another is "gliding", reflection in a line followed by translation along it. This painting, "Angels and Devils" by M. C. Escher (1898–1972), has symmetry elements decreasing in size

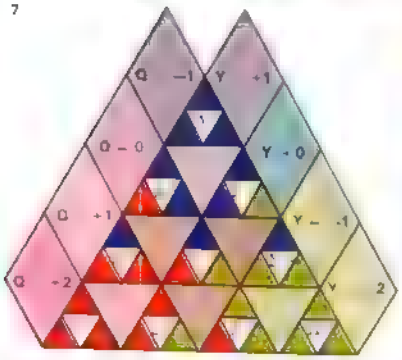
6 A

B



6 Why does a mirror turn an image right-to-left and not upside down? The answer to this confusing question is that the mirror's transformation is neither right-to-left nor up-to-down, but back-to-front. The left side stays on the left and the top stays on top, but the back becomes the front. Because hands (A), like other pairs of body parts related by mirror symmetry are called "right" and "left", a right hand becomes a left one, starting the confusion. Some lack of symmetry is not obvious. In man (B) the heart and other organs are on the right if he considers his mirror image

7



7 Symmetry considerations are basic to nuclear-particle physics. Many fundamental particles are now known, but the laws governing their occurrence and properties are poorly understood. One attractive theory has them composed of "quarks". The diagram shows how the d-quark (blue), the u-quark (red) and the s-quark (green) might combine in threes to form each of 10 particles called hyperons (−, 0, or +). The charge (Q) and "hypercharge" (Y) of each hyperon is correctly predicted by this type of symmetry classification, which also predicted the Ω⁻ particle before its discovery

8 A

8

8 One of the most paradoxical of physical laws is that of time symmetry: any process can go backwards. This may seem absurd but in a film of two or more billiard balls colliding (A), it might not be possible to tell if it was run backwards. The reversed film would still show a possible physical event. But it might not be a probable one and most processes (eg throwing a stone into a pond) are unlikely to reverse. Nonetheless, dispersed molecular motions could converge on a stone and eject it spontaneously through a calm surface. Similarly, light rays can always retrace their paths exactly, so a camera could be used backwards (B) as a projector, exchanging object and image, and remain in focus

The language of space: topology

When the comic film character Monsieur Hulot traces along a tangled hosepipe from a tap and finds it leads back to the tap again, why does the audience laugh? And what is so odd about a household hint from *The Times* of London: "Mending a hole in a tablecloth lay the cloth on a table with the hole upper- m ist [illegible]"? Both items offend our instincts about topology – a branch of mathematics that deals not with shape or size but with much more fundamental properties of objects and of space

Spheres, nets and knots

It is a topological truth that, regardless of its length or curvature, a hosepipe has two ends Similarly, we feel sure that no matter what the size of a tablecloth, or the outline of a hole in it, it would be hard to spread the cloth out with the hole underneath. Topology takes such intuitive matters and formalizes them into mathematical logic. It is concerned with all those properties of objects that are unaffected by any change of form, however extreme. For instance, any simple solid object without holes is a "ball" to a topologist, for if made of deformable clay it could be rounded into a ball without being torn. Thus topology is sometimes called "rubber sheet geometry".

A button with four holes in it is not a topological ball It is a "quadruply connected solid" because you would have to make four cuts in it, opening out the holes to the edge of the button, to make a shape that is a topological sphere. Small creatures living on the button would find it a different kind of space to the surface of a sphere. Any closed curve on a sphere [1], for example, must fence off and enclose a definite area On a button this is not so. A closed curve round one of the holes does no such thing. The pure topologist's main concern is to decide whether particular abstract entities (objects or spaces of many forms and dimensions) are or are not topologically equivalent Human intuition in comprehending the basic topology of even simple figures is relatively limited and sometimes leads to wrong conclusions

This strange branch of mathematics has links with the real world [7]. An electrical circuit is a topological entity, for example; its exact layout does not matter because only the pattern of interconnections is electrically significant. Graph theory [3, 4], the branch of topology that handles networks, is fundamental in advanced circuit design. And the age-old crafts of knitting and weaving are really exercises in applied topology A loop with a knot in it retains that knot however it is deformed and cannot be "undone"; it is topologically different from an unknotted loop. Textile manufacturers practise topology in their efforts to produce garments with specific topological properties, ones that can be knotted in one piece or that will not unravel if a fibre breaks [Key]

A mathematical playground

Most serious topology has, as yet, little to do with the practical world. No branch of it is as closely tied to human affairs as, say, arithmetic is to banking. It is therefore a subject full of potential – for time and time again in the history of mathematics such theorists' playgrounds have become the workshops of a new science or discipline. At present, however, theorems in topology, although proved

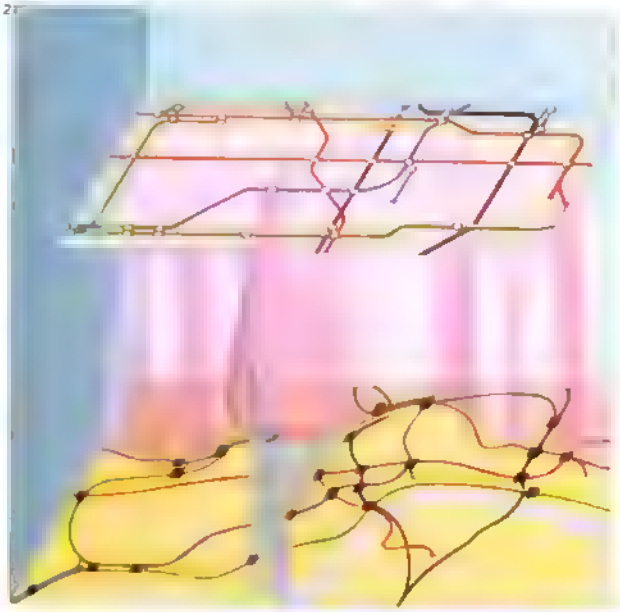
CONNECTIONS

See also

[illegible]

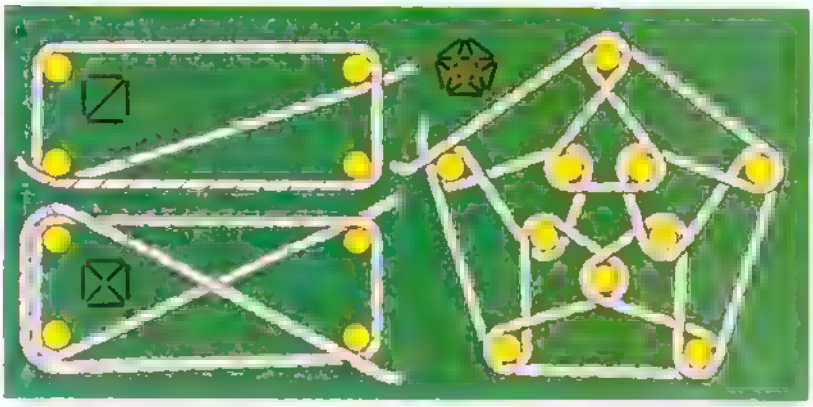
Maths and mapping

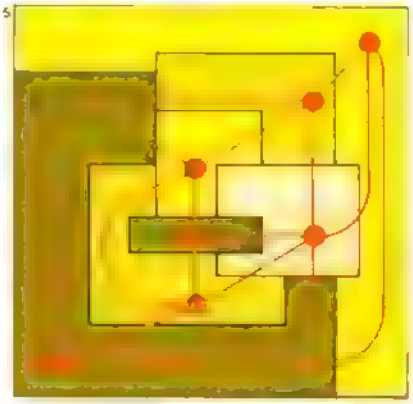
1 "Any closed curve divides a surface into an inside and an outside, and a line connecting these crosses the boundary an odd number of times." This theorem may seem obvious (A) but is not true of all surfaces. A table and statue (B) are topological balls on which the theorem holds. But a chair and man (because of his alimentary canal) have holes through them, like a doughnut (C), which need not obey the theorem



2 The London Underground map is a very distorted plan of the lines. But they correspond point-by point, and two points joined on the map are connected in reality This fundamental test makes them topologically identical

3 Königsberg, a Prussian town, posed a teaser that led to topological graph theory Could you take a stroll crossing each of its seven bridges only once? In 1734 the Swiss mathematician Leonard Euler analysed the problem to form a theory of the traversability of a network or graph shown superimposed on the city). One bridge has to be crossed twice The problem depends on connections, not distances





4 A graph traversable in one pass must have an even number of lines meeting at each junction. The pentagram C is traversable but the rope makes a double pass on rectangle B and begins and ends at different points on rectangle A.

5 A flat map needs no more than four colours to prevent adjacent areas sharing the same colour. This unproved theorem is part of graph theory, for every map can be drawn as a graph with areas as junctions and boundaries as lines

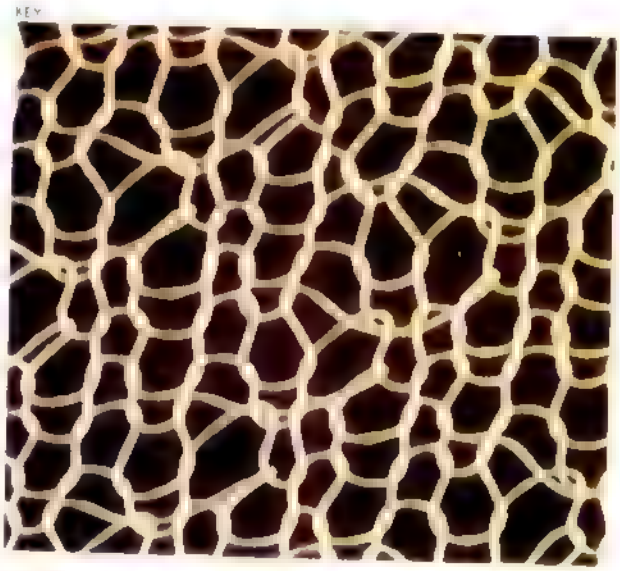
with full rigour, are less directly useful than those of a subject like geometry

A typical topological theorem says that in colouring a flat map no more than five colours are ever needed to ensure that adjacent areas need not share the same colour. The theorem does not state how this can be accomplished for any given case but merely asserts that it can always be done somehow In fact, four colours may be sufficient [5], although this has not been mathematically proved Similarly, it is topologically certain that however briskly a cup of tea is stirred, at any instant at least one point in the liquid is not moving. The topologist is not concerned to identify this point; he just proves that it must exist In different types of space different theorems hold good. There need be no fixed points in a stirred inner-tube full of water; on a doughnut, up to seven colours may be needed for a map without adjacent colours [6]; and on a Möbius strip, up to six colours may be needed

A Möbius strip is some sort of vindication of Monsieur Hulot for it is a contradiction of the strong human intuition that a piece of paper must have two sides. It is named after the German astronomer and mathematician August Möbius (1790–1868). His strip can be made simply by cutting out a ribbon of paper, making a half turn in the middle of it and sticking the ends together to form a twisted loop. This loop now has only one side, as you can prove by drawing along it with a pen, never going over the edge until you meet your starting-point again Cutting along the line creates another surprise

Twisted space

Topologists study such "twisted spaces" in more dimensions than two, hard though they are to imagine. Indeed it is topologically entirely possible that the universe itself has a Möbius twist in it. One result of this might be that a traveller who went far enough out into space would return reversed in mirror-image fashion with his heart in the right side of his chest. Glove manufacturers might then be able to make only left-handed gloves and ship half their output around the universe from where they would return as matching right-handed ones

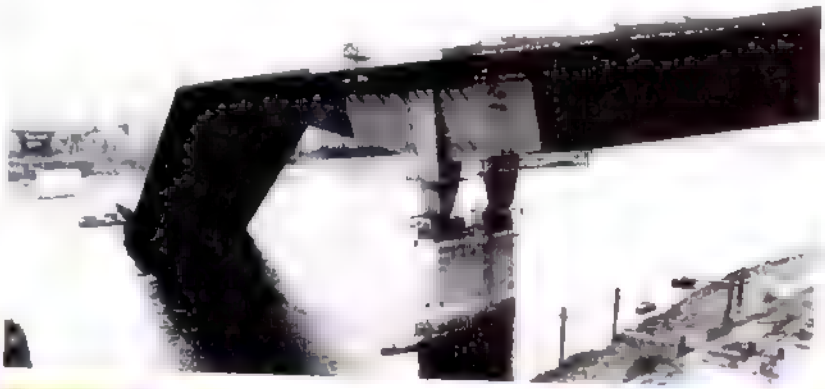


The significance of topology in textile manufacture is shown in the structure of a pair of tights photographed through a microscope The complex system of knots is designed to avoid a "run" in the garment if a fibre is broken

6 On a doughnut or torus a map can need up to seven colours to prevent adjacent areas sharing the same colour The map shown (with its mirrored reflection for completeness) needs all seven because each area touches the other six The sections form a continuous helix winding round twice before closing on itself

7 Any structure has many "modes of failure" Engineering disasters such as the collapse of the River Yarra bridge in Melbourne can occur because of modes which the designers have not recognized As a novel design such as the box girder bridge is refined and made lighter and cheaper, unsuspected modes of failure may be discovered the hard way Classification of abstract entities by their "spatial" properties can help here. Topological "catastrophe theory" is concerned with the ways shapes can change; its principal application is to morphogenesis, the biological study of the ways in which organs and tissues develop and form

7



6

8

8 A ball covered with fur cannot be smoothed down all over; at least two crowns must remain where the fur radiates from a point or piles up at one, or a set of partings. This "hairy ball" theory governs the way directions, like hairs, can be aligned on a sphere. If they are lines of magnetic flux, it shows that every magnet must have two poles. If they are wind directions on the globe, the theory proves that somewhere in the world the wind is not blowing

9

9 These ants, by M C Escher (1898–1972), illustrate the counter intuitive properties of the Möbius strip All are on the one side but appear to be on opposite sides A "strip" made with two half twists does have two sides. The number of twists dictates the number of sides and dramatically affects the result produced by cutting along the middle Topology allows us to explore and describe such spatial relationships

Maths and mapping

Can you read a map? A blueprint? A circuit diagram? Morse code? Then you are a mathematician because these are all examples of mathematical mappings. The idea is simple, a map is any way of relating one set of objects to another set [1]. In a geographical map [3], every one of the infinite number of points on the earth's surface corresponds to (or is "sent" to) just one of the infinite number on the map. Similarly blueprints and circuit diagrams map certain features of a physical object onto a pattern on paper

Maps and their meaning

One interesting thing about such maps is what they can and cannot do It is impossible for example, to map the whole globe onto flat paper without sacrificing some features to preserve others. True directions on a map are impossible to achieve without some distortions. But even distorted mapping is still mathematically acceptable – like the hidden painting of a skull [Key] The apparently meaningless set of distorted smears is an anamorphic painting (one that appears in proportion when viewed from a particular angle, generally using a lens or mirror), designed to map to a recognizable image when viewed correctly

Mathematical maps embrace much more than these simple correspondences of points in space. They deal with anything: points numbers, sets and abstract entities with no meaning beyond themselves. They even handle the mapping of a set of objects onto itself This seeming paradox is commonplace, for example, in secret coding. A code is a rule for replacing each letter of a message by another from the same alphabet, it is a complex mapping of the alphabet onto itself Similarly the two-times table relates a number to its double and (if we include fractions in it too) is a mapping of the set of all real numbers onto itself

One-to-one maps are not the only kind The "zero-times table" that takes all numbers to zero is a good mathematical map But the mapping must specify a definite image for every element in its "domain" of operation. Therefore there is the old puzzle of the village whose barber shaves everyone who does not shave himself. This purports to describe a mapping that sends all the "shavers" to themselves and all the "non shavers" to the barber It is not a well-formed mapping because it leaves the barber himself in a paradoxical position. Does he map to himself? (That is, does he shave himself?) If so, he shouldn't, and if not, he should

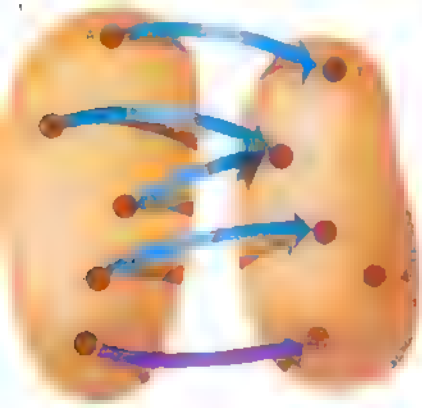
By contrast the marriage map between *n* men and *n* women (where *n* is any whole number) is a proper map and defines a possible set of marriages. Assuming that each individual has some order of preferences for his or her *n* possible partners, then it is a wry outcome of mapping theory that of all possible such maps – ways of pairing off the men and women – one and one only is stable. In all others cases will inevitably occur in which a couple, not married to each other, will prefer each other to their own spouses

Maps between one thing and another

The above mappings are examples of "discrete" instead of "continuous" maps Unlike those of points on a surface or a set of all real numbers they handle a finite set of elements only. The mapping of telephone subscribers

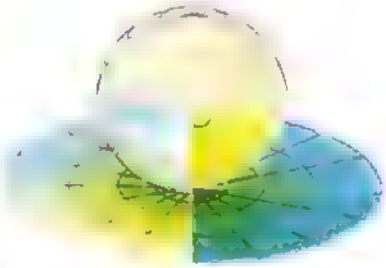
CONNECTIONS

See also

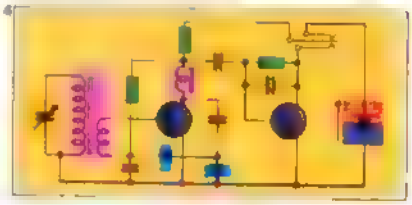


1 A mathematical map relates one set of objects (eg A, B, C, D) to an "image" set (eg 1, 2, 3, 4), symbolized by the arrows One-to-one correspondence is not necessary, both B and C are sent to 2, D is sent to 3 and nothing is sent to 4 But the map must act on every object in its domain Without the purple arrow E would be unimaged and the mapping therefore improper

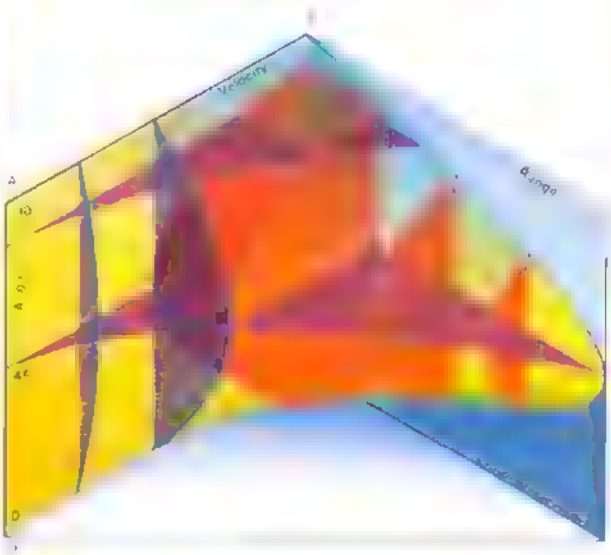
2 Light is transmitted by a fibre from one point to another the bundle maps an object into its image (A, it may be deformed by reduction (B or by scrambling the fibres (C) This is called fibre-optic mapping



3 Each point on the globe is sent to one on the map in this diagram. The "zenithal projection" is mapping the Southern Hemisphere, each point is projected along a line from the North Pole onto a plane touching the South Pole. The Equator becomes a circle as do other lines of latitude Lines of longitude become radii The scale is not constant it increases dramatically towards the edge of the map



4 An electronic circuit diagram is a map unconcerned with scale or shape It shows the connections between components using stylistic conventions – wires, for example, become straight-line segments



5 The muzzle velocity and elevation angle of a gun are mapped onto the range of the shell The plane OAB is a "map" of angle and velocity, the height of a point (P) represents elevation angle and its horizontal distance from OA represents muzzle velocity The perpendicular distance PS from the surface at P represents the range for those settings The surface contour shows that range increases with velocity and is greatest at 45° elevation

6 A reflex AM radio-receiver circuit has the circuit diagram shown in illustration 4 They do not look alike but are related by a mapping which ensures that the connections to each component are the same in both the physical and schematic layouts The manufacturer is unconstrained by circuit diagram conventions and routes the wiring for tight packing of components, for example But adjacent components may interact, by their electric or magnetic fields, through space

onto their telephone numbers is like this too. Each subscriber has his own number, but not all numbers are represented. For example there is no number 000 0000. A reservoir of possible but unused numbers is held by the telephone companies. This map illustrates a new point too. Our previous examples sent points to points, numbers to numbers, people to people, this one sends people to numbers.

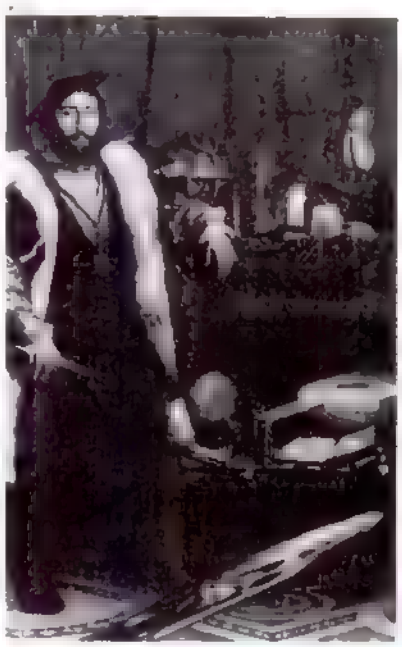
The mapping of one set of entities onto another apparently quite different set is a powerful mathematical technique. For instance, analytical geometry maps geometry onto algebra [11]. Each geometrical curve or line is sent to its corresponding equation, and for each geometrical theorem there is a corresponding algebraic identity. The mapping preserves the relational features of geometry so that geometrical problems of great difficulty (for example, those in many dimensions) can be mapped by easier-to-handle algebra, solved and then mapped back to give the required geometrical truth.

Morse code is another example of such mapping. Letters and numbers are mapped onto combinations of dots and dashes. These in turn can be transmitted as short and long flashes of light or pulses of electric current. At the receiving end the dots and dashes are mapped back to letters and numbers.

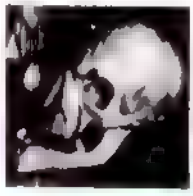
A vast range of scientific and technical enterprise depends on mapping from the real world into symbolic systems that preserve the important features. An astronomer maps the positions of the heavenly bodies into terms of a set of equations. From the ensuing calculations, he can recover terms that map back into future positions of the bodies and perhaps predict an eclipse.

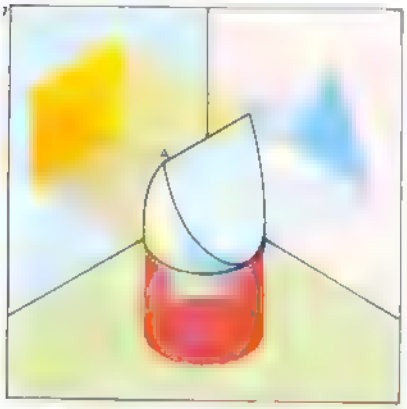
Mapping: theory into practice

It is the business of mathematical science to make sure that mappings work. All scientific theories are maps in this sense and so are the calculations and designs of an engineer who decides on paper that an aircraft as yet unbuilt can fly. Like geographical maps of the globe, these maps sacrifice some features to preserve others and so they are incomplete. They fail totally in some areas of experience and distort so badly in others that they are useless for any practical purposes.



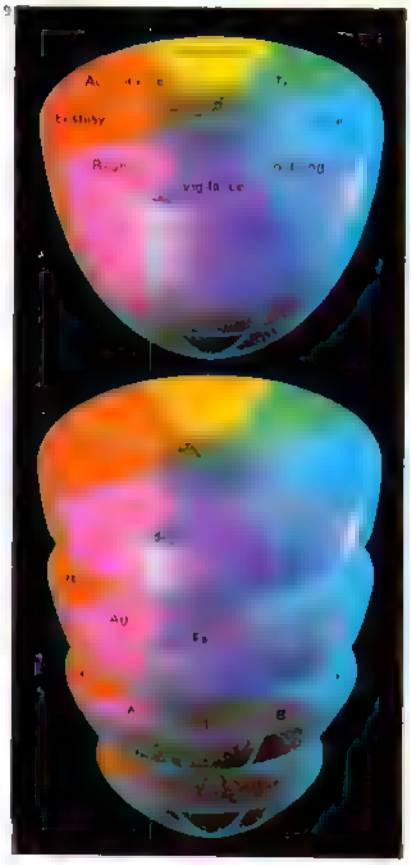
Anamorphic art is a technique in which an artist draws or paints a familiar shape in a grossly distorted form. It is an example of mapping, just as maps of the earth's curved surface can be drawn on a flat sheet of paper by a suitable choice of projection. In this painting part of "The Ambassadors" by Hans Holbein (1533) the stretched shape at the bottom is a "map" of a skull. It can be seen (below) by viewing from the lower right.





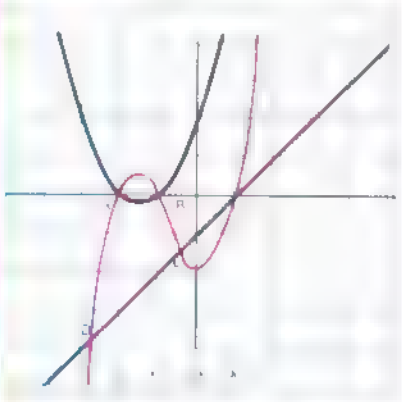
7 A draughtsman's projection is a mathematical map in which many points on the object go to the same point on the image. In the blue triangular projection all the points on the line AB are sent to the top vertex of the triangle. They retain their identity in the yellow square projection, but lose it in the circular red one. All projections compress information in this way, so draughting makes use of several

8 The male and female sex organs are both distortions of a primitive system and are mappings of each other in the mathematical sense. The testicles (A) map into the ovaries and the penis and bladder map into the elaboration of uterus and urogenital tract (C). This suggests that both systems may have evolved from a common one (B) and that, in nature, different structures can have the same form.



9 The psychologist Plutchik mapped the emotions onto this "emotional solid". The most intense ones map to the top. The next layer shows diluted versions of these (by their initial letters): delight, anger, expectancy, disgust, sadness etc, and the next layer happiness, annoyance, interest, boredom, pensiveness etc. The lower marks neutrality. Any sequence of emotions is a "worm track" through this solid.

10 A weight suspended by a spring in water (A) and an electrical circuit (B) are mathematical maps of each other. If disturbed, the weight will vibrate with decreasing vigour and an electrical pulse will cause oscillations in the circuit to fade away gradually. Both examples have an energy-storing element (the spring or capacitor), an inertial element (the weight or inductor) and an energy-dissipating element (the water or resistor).



11 "Mathematics is the art of saying the same thing in different words." (Bertrand Russell.) The curves are geometry, their equations algebra. The mapping between these is analytical geometry. The two curves intersect at A and B; their equations will yield two solutions whose values give the co-ordinates of A and B. Similarly the straight line intersects one of the curves at C, D and E whose co-ordinates are found by solving the equations simultaneously.

Facts and statistics

In the next minute at least 60 and not more than 310 babies will be born. This statistical claim requires no knowledge of any individual women. It just assumes the average world birth-rate of three per second – and it has only one chance in a thousand of turning out to be wrong

Making reliable statements about chance events is the business of statistics. Think of a tossed penny – the classic uncertainty. An unbiased penny tossed a million times gives, with 99 per cent certainty, between 498,700 and 502,300 heads and the rest tails. Conversely, if bias is suspected, one toss will not confirm it. But a million tosses giving 500,000 heads indicates, again with 99 per cent certainty, that the bias of the coin used is between 0 4987 and 0 5013 (a perfect coin has a bias of 0 5000)

Chance and certainty

Phrases such as "99 per cent certainty" are common to all reliable statistical statements. Certainty is never 100 per cent and for this reason a reputable statistician always states his error and confidence limits. Ninety-nine times out of 100 he would be right in bracketing the penny's bias between the given limits. Only once in 100 times would a coin of greater bias give, by chance, equal numbers of heads and tails. If it is really necessary to bracket the penny's bias with greater confidence, it would have to be tossed more times In statistics there is always this trade-off between the information necessary and the reliability of the knowledge it yields, with complete certainty forever unattainable The art of practical statistics lies in knowing the probability that suffices for the task in hand and knowing what is the sufficient amount of data to collect in order to derive it

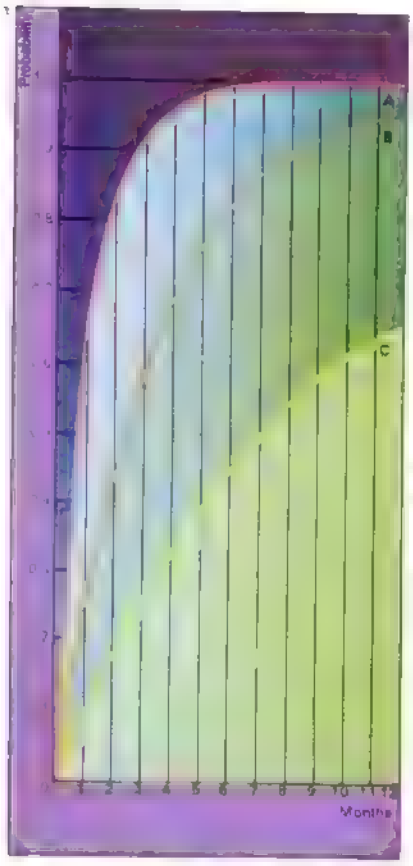
Making good and bad guesses

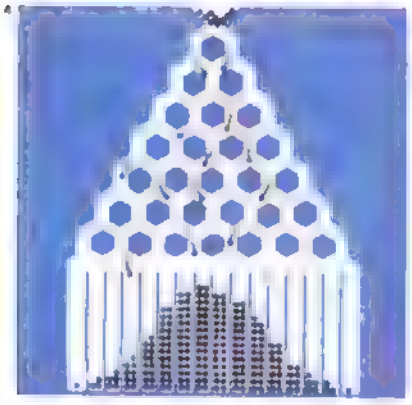
Insurance companies depend on statistics Will a client aged 20 die in 40 years time? Nobody knows. But an insurance company, with its records of thousands of men and women, estimates how many clients are likely to die and be the subject of claims and therefore how much it must charge in premiums to keep in business. From its widely amassed survival statistics it can deduce how dangerous overweight, smoking and so on, are to health. This is achieved by seeking "bias" in the death records of various groups, just as one might seek it in the tossing records of various pennies

Medical science gains from this too. It was statistical analysis that correlated the taking of the drug thalidomide during pregnancy with deformed babies and cigarette smoking with lung cancer. But such correlations need careful interpretation. Statistics cannot say why smokers are more likely to contract cancer than non-smokers. Perhaps people predisposed to lung cancer also tend to have a taste for smoking – an odd hypothesis, perhaps, but one that is statistically feasible Similarly cancer of the cervix in women shows a slight but definite correlation with the number of children they have borne Does this mean that childbirth causes cancer? Further studies show that the correlation fails for Jewish women. This clue leads to the conclusion that the correlation is with sexual activity rather than with its natural outcome In fact the correlation arises because of the irritant substances that can form under the

CONNECTIONS

See also
Kids and probability



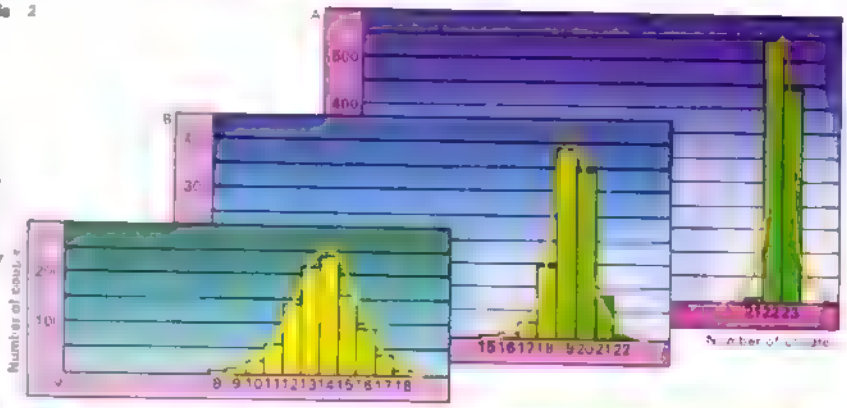


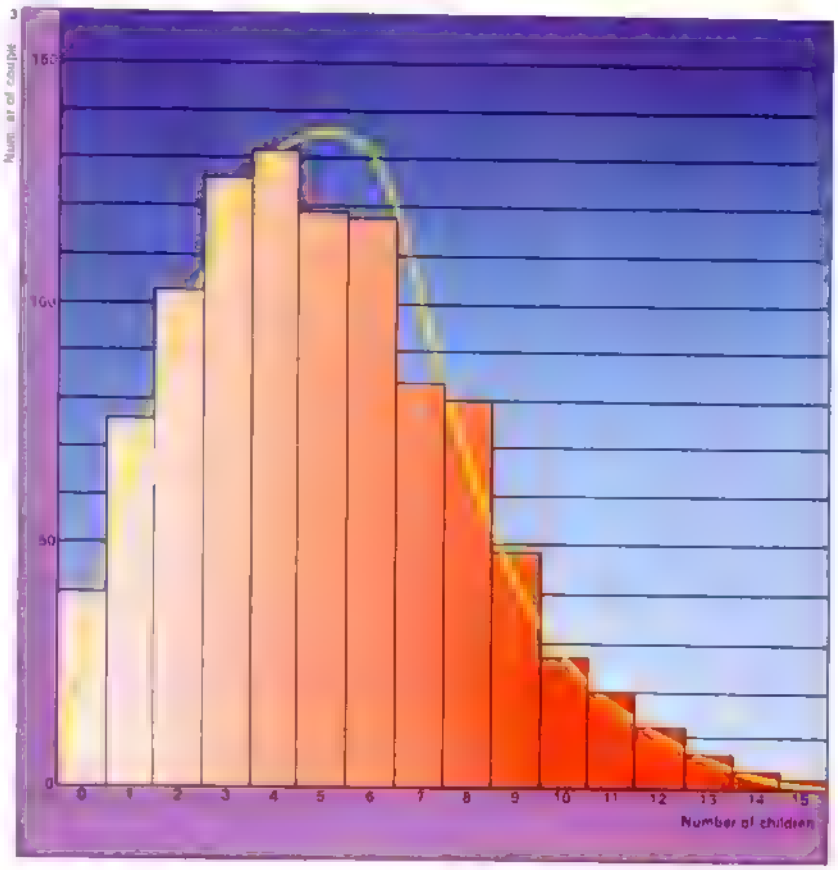
1 Conception depends on many small chances, even when intercourse occurs during ovulation A fertile couple s chance of conceiving might be as curve A, reaching 0 9 (90%) in 3 months. An average couple (B) might have a 60% chance of conceiving in this time and one of low fertility (C, of only 25%, with only 60% chance of conceiving in a year.

2 Computer calculations show the range of family size 1,000 couples of each type might expect after 25 years A will probably have 21–23 children; B 15–21 and C probably 10–17.

3 Even in a primitive community very large families are rare. Limiting factors include the death of the mother or her becoming infertile If, for such reasons, the chances of a family being complete after the birth of the first child is assessed as 4%, at 8% after the birth of the second child and so on, then families of high, average and low fertility each have an average of five or six children and produce the same curve, shown here.

4 A rain of balls through this Galton board is distributed in the bell-shaped "normal error curve. This and similar curves are commonplace in statistics It shows the outcome of events under many individual chances most stay near the average a few stray farther away



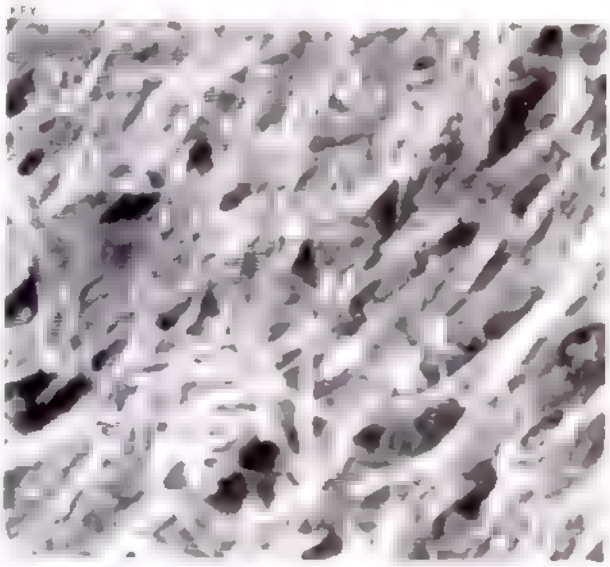


foreskin of an uncircumcised man if he is careless about personal hygiene. Jewish men, being circumcised, do not expose their wives to this slight hazard. So mathematicians must not jump to hasty conclusions. Correlations are not causes (merely clues), and statistical data are dangerously easy to misuse.

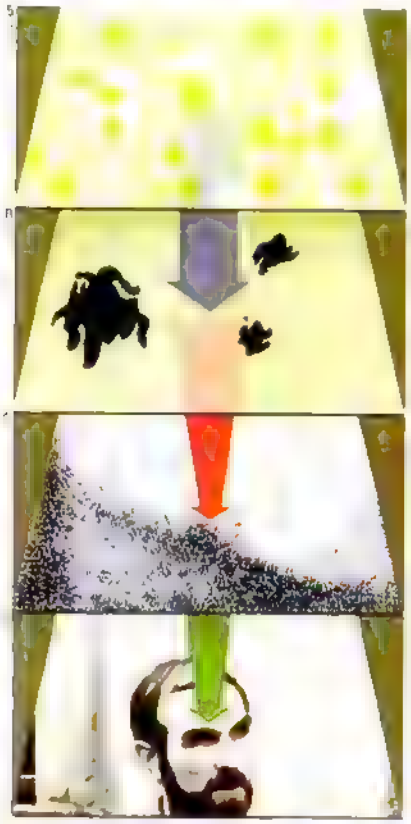
Molecules and magnetic tape

In a sense the whole world is ruled by statistics for its individual atoms and molecules are, by the uncertainty principle, not completely predictable. Only when considered in countless millions is their behaviour reliable. It is most unlikely, for example, although theoretically possible, that all the air molecules around someone would chance to rush away spontaneously and leave him to suffocate. On a smaller scale chance molecular fluctuations are inevitable and modern technology, in its quest for sensitivity, occasionally encounters them. A good audio amplifier with the volume turned up, for example, produces a slight hissing which is the amplified random motion of the electrons in the input circuit. It is as if the amplifier handles information by "tossing electrons" and residual uncertainties cannot be avoided.

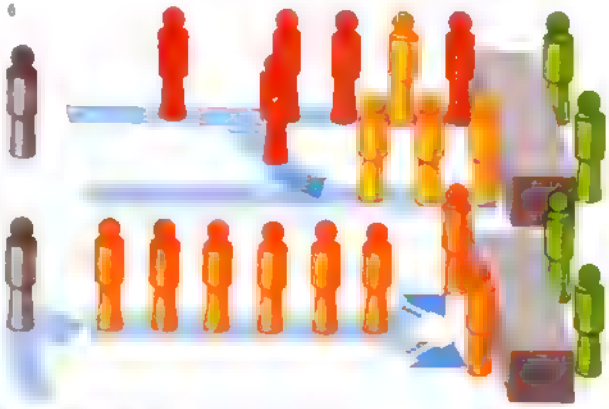
In a similar way magnetic recording maps an audio signal onto millions of metal-oxide particles on the tape [Key]. Each can have one of just two magnetic states, equivalent to heads and tails. The faster the tape runs, and the wider the track processed by the recording head, the more particles are used to record a given sound by the changes in their distribution and magnetic states. For this reason the best quality machines use high speeds (38cm/sec [15in/sec]) and wide tracks (up to 1.26cm [0.5in]) to reduce tape-hiss to the minimum. Domestic recorders use lower speeds down to only 3.75cm/sec (1.5in/sec) and track widths down to 0.05cm (0.02in). They suffer accordingly from the smaller sample of magnetic particles from which they must reconstruct the signal. The same statistical principles underlie extraction of information from the tossing of pennies, the fate of smokers and the reading of magnetic tape – and make it highly likely that you will be able to draw your next breath.



Metallic particles on recording tape here highly magnified, have their magnetic state changed by the recording process. The quality of the recording depends on how many are affected.



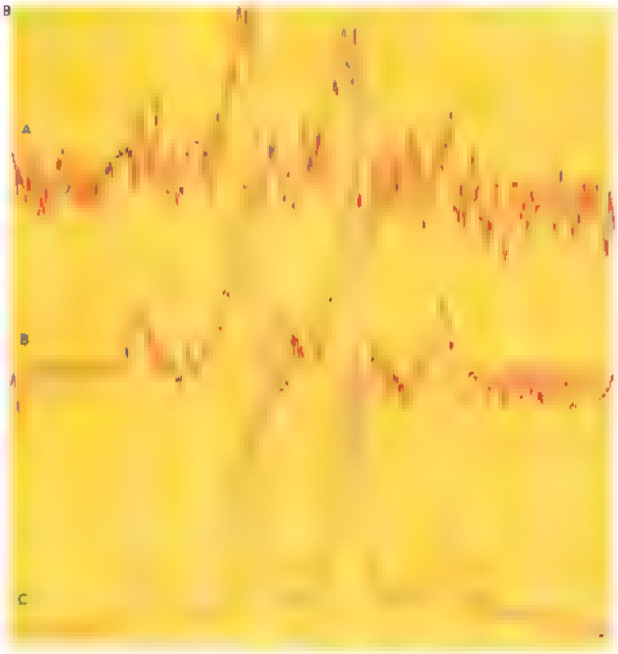
5 In photographic film light-sensitive grains are distributed in gelatine. Two photons (light particles) must hit a grain (A) to render it developable (B). In a random hail of photons this is pure chance for any one grain. But there are so many (C) that statistically the number of developed grains follows the illumination closely (D). There is a remote chance that the picture might look like something completely different.



6 People arrive at a counter quite unpredictably. How many clerks are needed for an efficient service? This is a question for queuing theory, which predicts that one queue served by two clerks is more efficient than two queues served by one each. It also decides how much switchgear a telephone exchange needs to install in order to handle randomly arriving calls, how many machinists a machine-shop needs to employ in order to cope with irregular repairs, and so on. Queuing theory can be used to prove that if a queue is not to grow infinitely long the person or machine serving it must be idle some of the time – a fact that some managers hate to face!

7 A person treading on a step removes a certain amount of stone from his path. Over the years the stone wears away according to the average distribution of paths down the steps. Since most people aim to go down the middle, but deviate randomly to either side, the steps tend to wear into the bell-shaped "normal error" curve. This is a statistical curve that, in time, draws itself.

9 A liquid of one colour is decanted into another and left. Soon the boundary becomes fuzzy and in due course the mixture is uniform. Any one molecule wanders at random, but the statistical effect of all their travels is a perfect mix, the most disordered arrangement possible. The gradual but inevitable increase of disorder is a basic law of physics.



8 Many scientific instruments must register a weak signal against a background of random interference. One strategy is to keep repeating the measurement. The signal is always there, the interference is positive or negative and statistically tends to cancel out. The top trace (A), from a nuclear magnetic resonance spectrometer, shows a spectrum heavily degraded by random interference. In B 16 scans have been added and in C 256 scans.

Odds and probability

A businessman worried by the prevalence of aircraft sabotage consulted a mathematician. "Don't worry," he was told, "there is only one chance in a thousand of a plane having a bomb on board." "But I do such a lot of flying," said the businessman. "Then always carry a bomb yourself," came the reply, "because there's only one chance in a million of a plane having two bombs on it!"

How to find the probabilities

This is an elementary but popular fallacy about probability theory. If two independent events each have a known probability such as one-thousandth, the chance of their both occurring together is indeed obtained by multiplying the two probabilities giving in this example one-millionth. But they must be independent: the chance of one cannot be altered by tampering with that of the other – such as ensuring its certainty.

This multiplication rule is one of the two great pillars of probability theory. The other, the addition rule, says that given two mutually exclusive events (such as rolling a one or a two with a die – both cannot be rolled), then the chance of either occurring is the sum of their probabilities. In this case each has a 1/6 probability, so if either one or two wins, the chance of success is 1/6 + 1/6 = 1/3.

These two rules, carefully used, can solve most problems of probability. They rest on a subtle sort of probabilistic "atomic theory" that takes any chance event as being compounded from a set of basic "equiprobable events". By calculating what combination of these will result in the desired chance coming up, its probability is obtained. But the notion requires subtle handling. Many misleading arguments depend on a deceptive choice of basic equiprobabilities. What is the chance of there being monkeys on Mars, for example? Either there are or there are not – and it could be argued that, since nobody has yet been to Mars, these mutually exclusive situations are equally probable. Then each has half a chance of truth and there is a 50 per cent chance that there are monkeys on Mars.

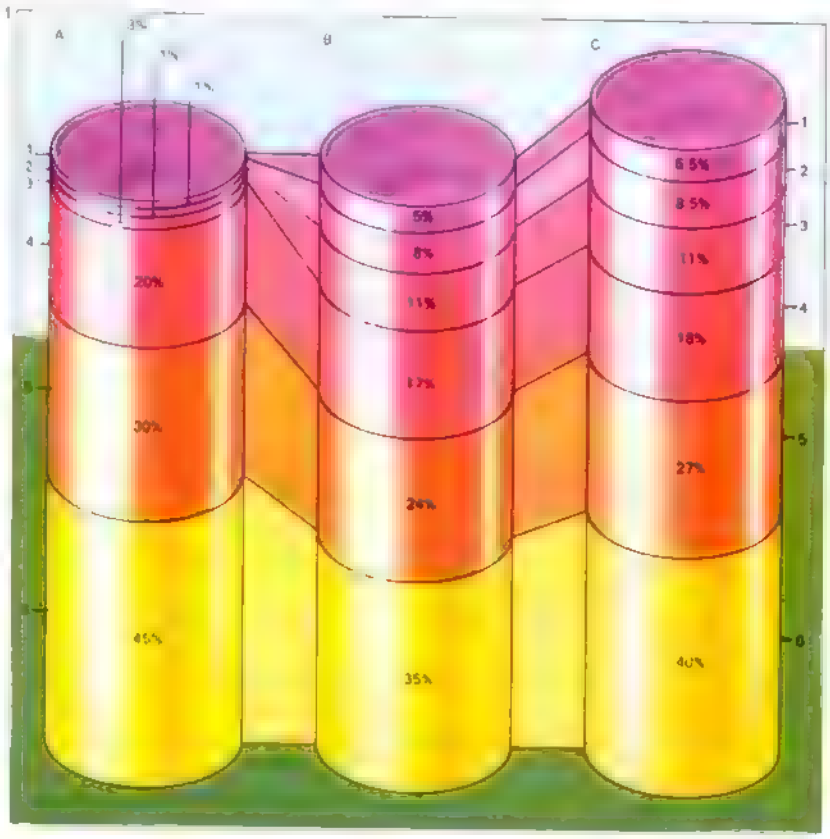
More subtly, what is the chance of getting one head and one tail on two tosses of a coin? It might be reasoned that there are only three basic possibilities: two heads, head and tail, and two tails. Only one of these is favourable, so the chance is 1/3. But this is not so. There are actually four "atomic" equiprobabilities: HH, HT, TH and TT (where H stands for heads and T stands for tails), of which two are favourable. The chance is 2/4, or one half.

Calculating the chances of success

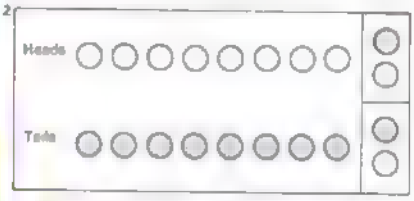
In mathematical notation, chances vary from 0 (impossible) to 1 (certain). If there are 7 equiprobable possibilities, and 2 of them will result in success, the chance of success is 2 in 7, or 2/7, or 0.2857. This can also be expressed as 28.57 per cent, or in betting parlance 2 to 5 on, or 5 to 2 against. Such figures make most intuitive sense when applied to situations that can occur many times. In a run of 7,000 trials each with a 2/7 chance of success, about 2,000 successes would be expected. A gambler would break even in the long run by accepting odds of 7 to 2 (that is £7 return for a £2 stake). Where the basic equiprobable events are clear and knowable (as in the fall of coins, dice or cards), probability theory can give unambiguous chances of success for any outcome. All casinos and

CONNECTIONS

See also
Facts & statistics



1 The bookmaker aims to offer odds that give him the same predictable profits whichever horse wins. Thus if he received £3 on one horse and £5 on another, he might offer odds of 7/3 on the first (4 to 3 on) and 7/5 (12 to 5 on) on the second. Whichever wins, he pays out £7 and makes £1 profit. So his odds reflect the money bet. "Outsiders" attract little money, so he offers long odds on them. The chances of six horses are shown in A. 1, 2, 3 are outsiders with very long odds against them. But novice betters find them unreasonably seductive, so the money placed distributes itself as in B. The bookmaker changes the total odds upwards in his own favour, as in C. He is sure of a profit in the ratio of C to B. But even so, some winning odds – on the favourite (6) – are undervalued in C compared to the "reality" of A: 40% (he offering a return of 100 to 40) compared to its actual chance of winning, 45%. Hence 6 favours the punter and a series of such bets should clear an average of 10% profit to him. But the gullible backers of outsiders, in the long run, also lose. The same mathematical calculation of odds – probabilities – occurs throughout science. In atomic theory, for example, the location of an electron with in an atom is defined in terms of probabilities.



2 A tossed coin can land either "heads" or "tails". On each toss the probability of a head (or tail) is 1/2 (0.5) – the chances are evens. If a coin lands heads (or tails) eight times in succession, a gambler might be tempted to expect that a tail (or head) is more likely to occur on the ninth toss. But the mathematical probability of either outcome is still exactly 1/2 – an even chance.

3 Crown and anchor uses three dice inscribed with the six symbols of the matrix below, which shows all outcomes for the first die. "diamond" (five other matrices are similar). Players bet on their symbols against a banker, who returns twice the stake for one symbol displayed, three times for double and four times for triple. Assume each symbol is backed, giving six stakes "input" per throw. 20 out of 36 times (by the matrix), three different symbols come up and the banker makes no gain, returning three 2-fold stakes to the winners. On doubles (15 in 36) he pays out two on the singlet and three on the doublet and keeps one. On the only triple, he pays four and keeps two. So in 36 rounds he has gained (with one unit staked on each symbol) 15 + 2 = 17 of the 216 stakes. 7.9% return.

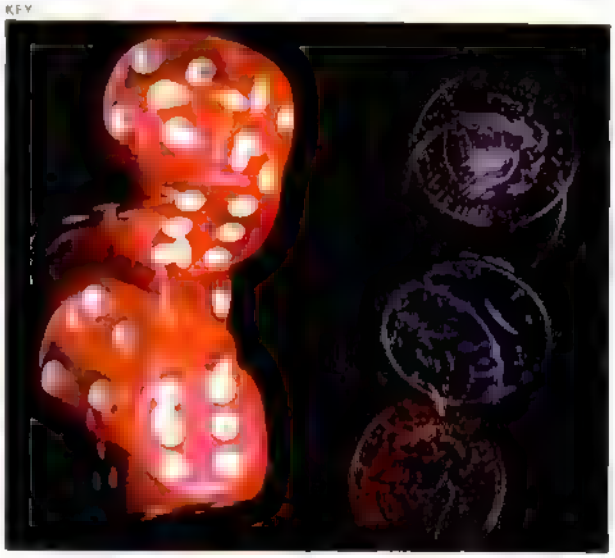
gambling houses use this principle to set fixed odds that give them a small advantage

In sports and business assessments, odds are subjective and different people guess them differently. By betting on the favourite in a horse-race with a number of unproven "outsiders", however, the gambler's chances of winning are demonstrably better [1]. If one of the horses is known to be doped, or a rival's business strategy is known, it is possible to place investments with better-than-average insight. This is the province of "game theory" – the theory of competing for gains against opponents who possess assumed aims and knowledge

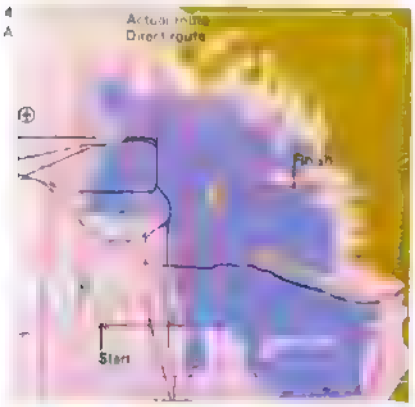
In the child's game of button-button a button is hidden in one hand and the opponent has to guess which He wins a penny if he is correct and loses one if he is wrong. What is the best strategy for the holder? If the same hand is always played, or hands are switched regularly, the opponent will soon outguess the holder. Game theory proves that the best strategy is to decide the switch at random, for example by tossing a coin before each round This is entirely foolproof; even if the opponent discovers the strategy he cannot win more than he loses in the long run But if two pennies are lost for a right-hand disclosure and only one penny for a left-hand one, the opponent could then win steadily by always choosing the right hand, and making on average bigger gains than losses. For this modification, game theory prescribes for the holder "weighted random switch" of 2:1 towards the left – say by tossing a die and playing to the right on 1 and 2, but to the left on 3, 4, 5 and 6

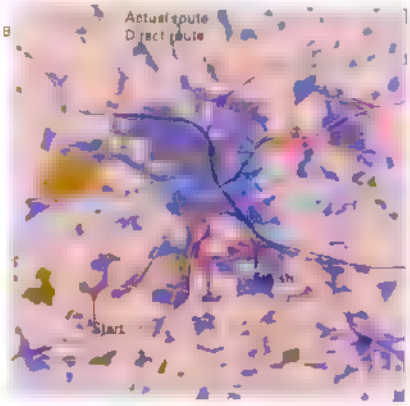
Uses in real-life conflicts

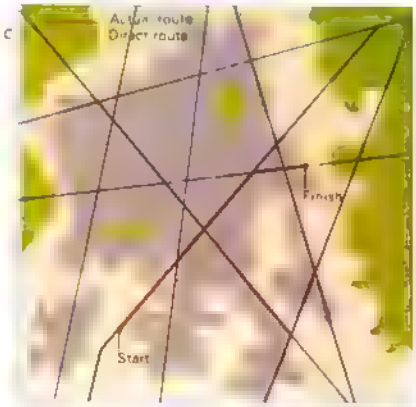
In real-life conflicts such as war and business, game theory is often used for clarifying options, but seldom slavishly followed If two people make an agreement, for example, game theory recommends to each that he double-crosses the other, for he will gain more if the other is honest. And in a world of unique events that either happen or do not, the whole concept of probability needs careful handling. Be warned by Peter Sellers's parody of a politician, who "does not consider present conditions likely"



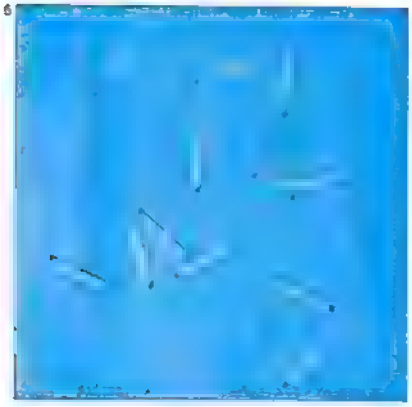
Probability theory cannot predict the outcome of a chance event such as the rolling of a die or the tossing of a coin. But in the long term (thousands of rolls) any one number on a die will occur with a probability of 1/6 (0 16666)



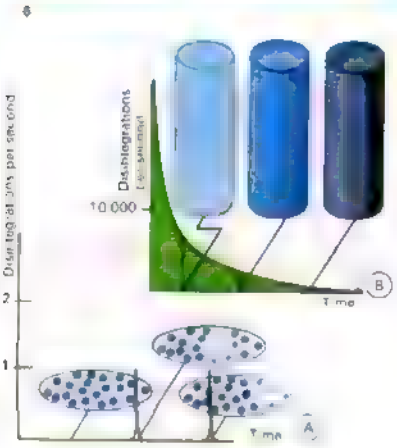




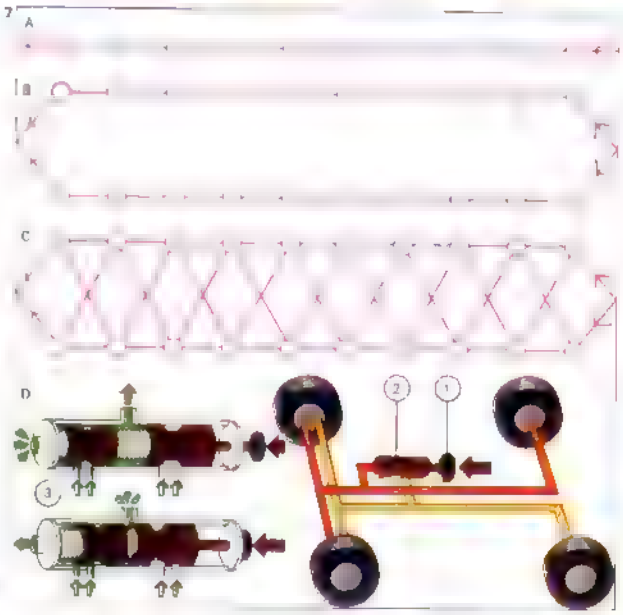
4 Is the rational grid layout of Salt Lake City [A] more efficient than the rambling European city of Cracow [B]? A diagonal journey on a grid forces you to traverse the equivalent of two sides of a triangle, even if you zigzag Probability theory shows that to facilitate many unpredictable point-to-point journeys, a random distribution of straight lines is best [C], a style close to Cracow's



5 In the "buffon needle problem", a match is thrown at random on a striped cloth. If the stripes are n match-lengths wide, the chance of it coming to rest across a line is $2/n\pi$. It is surprising to find π in this answer it enters because the match can lie at any angle, like a spoke in a wheel thrown onto the stripes. Mathematicians have evaluated π experimentally by repeated throwing



6 Radioactive atoms have a certain chance of decaying which is not affected by the presence or absence of other atoms. The rate of disintegration of a radioactive material is proportional to the amount of radioactive substance present multiplied by the chance that an atom will decay Whilst we cannot say when a single radioactive atom will decay [A], if ever, we can forecast the average behaviour of a number of radioactive atoms [B] Electrons, gamma rays or helium nuclei can be emitted.



7 A chain of components, all of which must work if the system is to function, is less reliable than its members. With 10 elements each of 99% reliability [A], the whole thing has about 90% reliability One improvement is duplication [B], with two such chains in parallel, the chance of at least one of them working is 95% But it is better to parallel each element separately, so that a paralleled pair will still function if either of its members is not working [C] Then each pair has a reliability of 99 9% and the whole chain of them has 99 9% The principle is built into a car's dual braking system [D] Pressing the pedal [1] moves pistons in the master cylinder [2] Three brakes work even with a leak [3]

8 A fly has an instinctive system of evading predators using an aerobatic pattern. As game theory recommends, it keeps making random alterations of course at random times Its course is then safely unpredictable, even to itself

The scale of the universe

Every object around us – indeed, all matter – is made up of countless tiny fragments called atoms. And the Earth is but a tiny speck in the vastness of the universe [Key]. But how large are these fragments? How big is an atom? And how large is the universe?

According to current thinking, the observable universe is about a billion billion billion (10^{36}) times as large as a single atom. But this statement gives no clue to the absolute size of either of them. To define the sizes of atoms, galaxies and the universe – as for a table or a garden – scientists use a series of units. An understanding of these is essential to a proper understanding of modern science – and to helping the imagination to grasp the range [1] between the immensity of the universe and the smallness of an atom.

Units of scale

Small objects can be measured in millimetres (about 0.04 inch) and longer distances are quoted in kilometres (about 0.621 mile). It is difficult to imagine the number of millimetres in a kilometre. But 10mm = 1cm, 100cm = 1m; 1,000m = 1km. Or, writing the numbers as powers of ten, 10^1mm = 1cm; 10^2cm = 1m; and 10^3m = 1km. Therefore one million, or 10^6mm = 1km. To denote something smaller a negative index is used: 10^{-1}cm (a tenth of a centimetre) = 1mm.

Today an atom is visualized as being almost all empty space with a few tiny sub-atomic particles near the centre which are surrounded by electrons. Very roughly, a sub-atomic particle [2] may be thought of as having a diameter of 10^{-13}cm. Ten billion (10^{13}) of them stretched out in a row might extend through a centimetre. The nucleus of an atom is made up of such particles – protons and neutrons – and may be 10^{-12}cm in diameter. An atom is the next jump in size: measured by pioneers of X-ray crystallography in ångström units, Å = 10^{-8}cm, an atom is about 100 thousand times as large as a proton. Atoms can be bound together to form molecules that can be grouped to make a volume of any size, molecules of gas; a crystal, a droplet of liquid, or all the water in the oceans. The paper of this page is about a few million atoms thick.

The wavelength of visible light is 4×10^{-5}cm to 7.2×10^{-5}cm. As a result, particles with a larger diameter than this can be seen using an ordinary microscope. To make smaller objects visible scientists use electron microscopes, because fast electrons have much shorter wavelengths. The smallest living organisms, such as bacteria, are microscopic. Smaller bodies such as viruses [3], which are submicroscopic, cannot live and develop alone but are parasitic on the cells of living organisms. All visible living things are made up of many millions of atoms.

Distances – from men to the stars

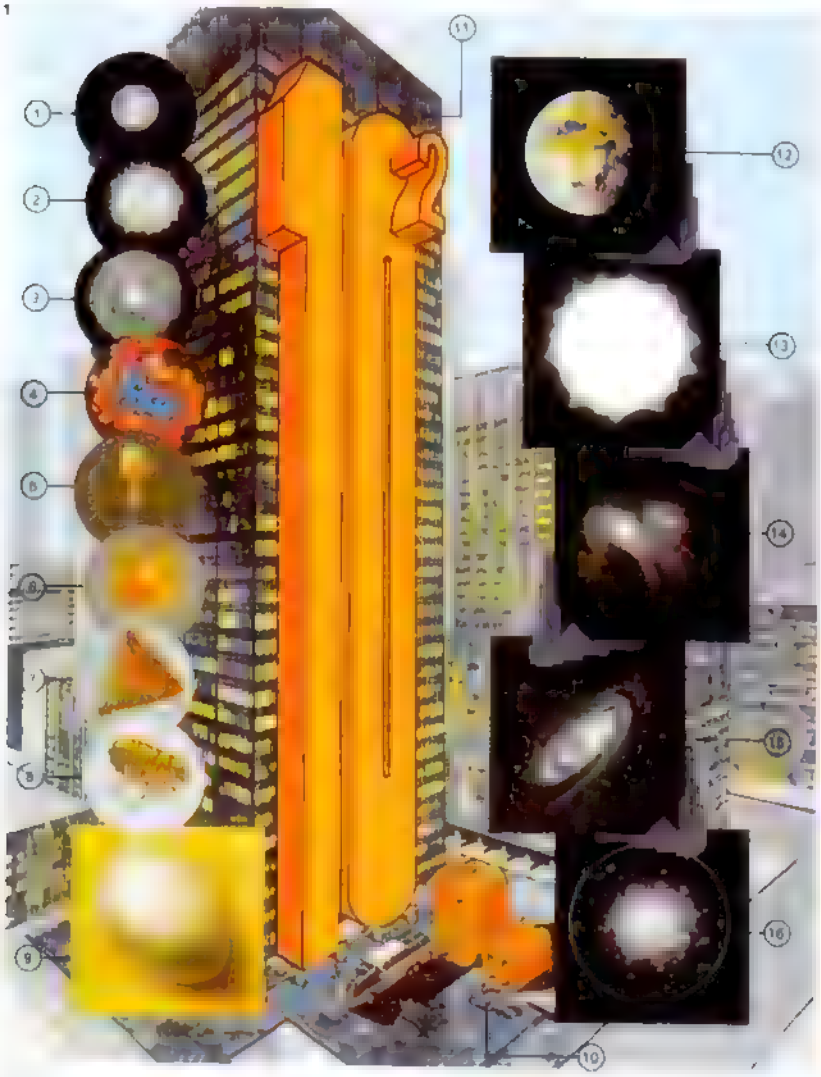
The tallest men are about 2m (6.5ft) in height and the Earth is more than 12,000km in diameter. The diameter of the Sun is more than a million kilometres. The nearest heavenly body to Earth is the Moon, about 384,000km away. Since man landed on the Moon and looked back to the Earth [7] this distance has acquired a more tangible reality.

The Sun is about 1.5×10^8 (150 million) km away from the Earth and the planet Pluto nearly 6×10^9km. These numbers are already becoming difficult to visualize and the whole

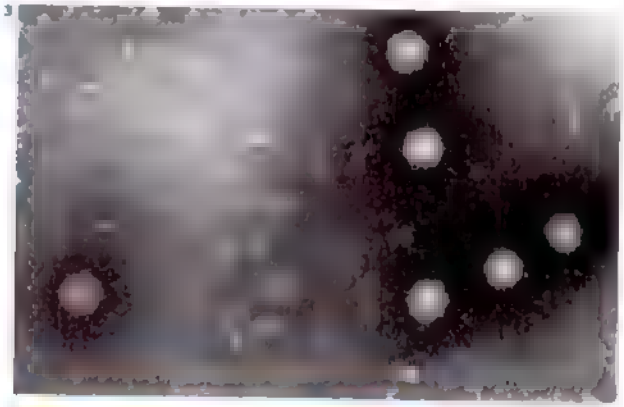
CONNECTIONS

See also

1 Within the known universe the dimensions of tiny, sub-atomic particles and the distance attainable by astronomers' telescopes stand in a ratio of about $1:10^{40}$. The objects shown spanning this staggering range are a proton [1], an atomic nucleus [2]; an atom [3], a giant molecule [4], a virus [5], a small cell: an amoeba [6]; a large cell, a diatom [7], a flea [8], a hen's egg [9], a man [10] 2m (6.5ft) high and one of his buildings [11], the Earth [12], a giant star [13], an interstellar gas cloud or nebula [14], our Galaxy [15] and the limits of the theoretically observable universe [16]. The 10m symbol is 10m (32.5ft) tall – five times as tall as a man – and the skyscraper [11] is more than 10^2 = 100m (325ft) tall, enough to dwarf the men and the bus.



2 Particles are so remote from our experience that indirect methods must be used to make them visible, such as the cloud chamber invented by Charles Wilson in the 1920s and developed by Patrick Blackett. It uses water vapour condensing on ions to reveal particle tracks and therefore identify them. Particles can also be tracked in bubble or spark chambers, or by using stacks of special photographic plates.



3 Viruses are too small to be seen under an ordinary microscope but can be seen by using an electron microscope. They are non-living matter but affect the properties of living cells.

4 The unique pattern of a finger print exemplifies the enormous number of individual cells that make up even the smallest piece of living matter visible to the naked eye. There are about 10 million cells in each cm^2 of skin.

Solar System the Sun with its planets and smaller bodies [8], is the merest speck in space Therefore to describe the geometry of the stars a different unit of distance is used the light-year, or the distance light travels in a year In just one second light travels 3×10^5 (300,000) km In a year light travels about 10^{13} km and the nearest star is more than four light-years from the Sun Another common unit is the parsec, equal to 3·26 light-years

The atomic nucleus is formed of densely packed particles, but the atom is almost all empty space Similarly, in the universe the atoms on the planets and stars are closely packed to form solids, liquids or gases But the stars are separated by huge distances compared with their diameters and so the universe around us, like the atoms of which we are made, is nearly all empty space

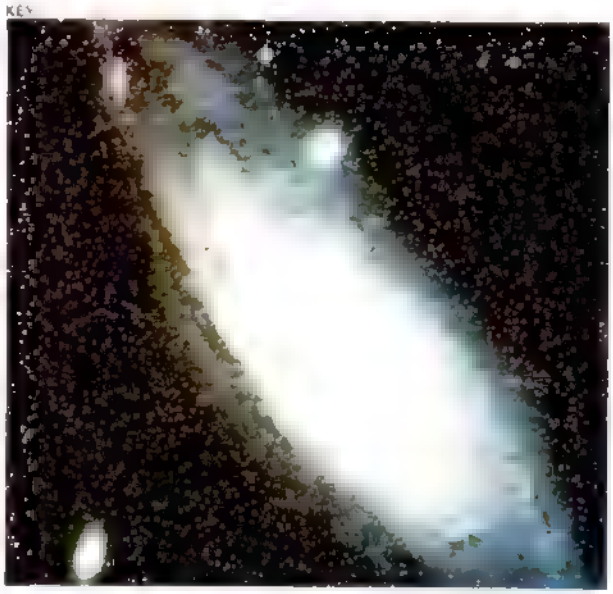
Towards infinity

The stars themselves, at their immense distances are grouped into great systems called galaxies or extragalactic nebulae So huge are the numbers of stars that through telescopes the galaxies look like great white clouds In our Galaxy alone, the Milky Way, there are more than 10^{11} stars. The Galaxy is about 10^4 light-years thick in the middle and 10^5 light-years across

Galaxies have greatly differing appearances according to the wavelengths of the light used to observe them One of the most rapidly developing branches of astronomy is radio astronomy in which observation is made by radio waves The largest known object in the universe is a galaxy designated 3C236, observed by radio telescope. A curious object in which the principal radio sources are two bulges at each end, it measures about 2×10^7 (20 million) light-years from end to end. The nearest large galaxy to our own is 2×10^6 light-years away The farthest objects probably lie at a distance of about 10^{10} light-years

The scale of the known universe from the atomic nucleus to the farthest star is about 10^{40}. But what is beyond the limits of present-day astronomy? Does the universe stretch to infinity? Or could it be curved in a curious way so that the nebulae that seem farthest away are in reality near to our own?

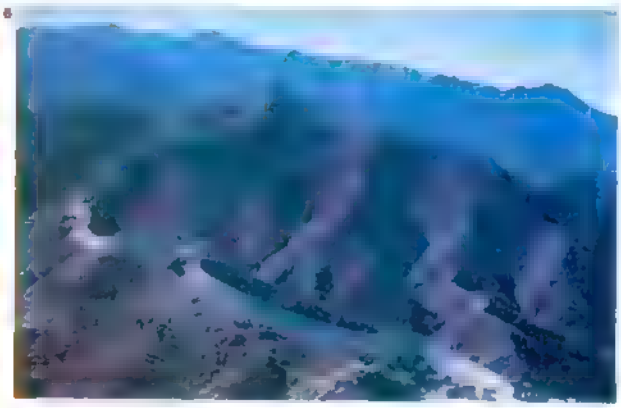
KEY



The spiral galaxy in Andromeda has a measurable size, but it is difficult to grasp 120,000 light years or 10^{18} km across.

5

6



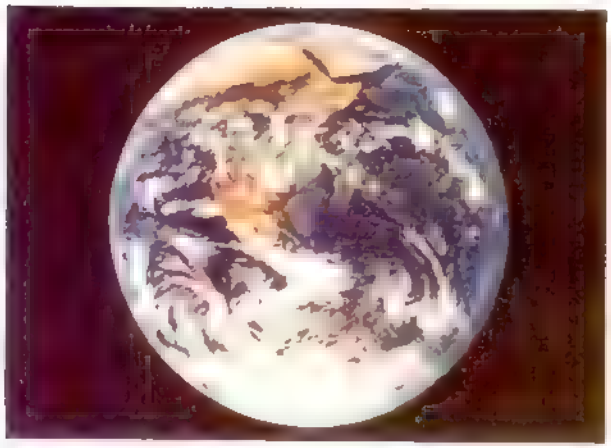
5 "Man is the measure of all things" said Protagoras in the 5th century BC – a humanist view that established man at the centre of the universe and related objects to a human scale. The person ification of nature was a theme of such Renaissance painters as Botticelli (1444–1510), who painted "The Birth of Venus" Modern science uses other units and man is no longer the standard of length

6 Man's most visible artefact, the Great Wall of China, runs for more than 2,400 km (1,500 miles) about five per cent of the circumference of the Earth It can be seen from well out in space

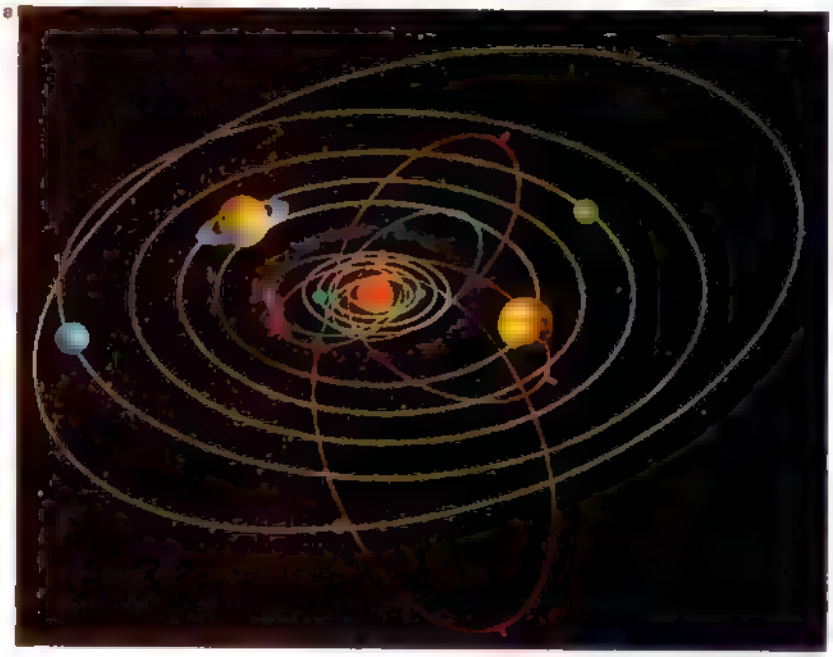
7 The Earth, as a body floating in space has acquired a new reality since man left his own planet and looked back at his terrestrial home For the first time, it has been seen as merely one, minute, heavenly body in immeasurable space The idea of other planets having other life forms is no longer regarded as improbable and current estimates put the number of possibly inhabited planets at many millions

8 The Solar System has been brought within man's reach by interplanetary rockets But beyond it the galaxies stretch endlessly the nearest to Earth a 2 million million million km away

7



8



What is an atom?

The first recorded suggestion that matter might consist of separate particles was made in the fifth century BC probably by Leucippus of Miletus [1] and the idea was developed by his pupil Democritus, who adopted the word *atomos* (from the Greek word meaning indivisible). John Dalton (1766–1844) revived the word at the beginning of the nineteenth century when he provided a scientific basis for the simple Greek idea. To Dalton an atom was a tiny indivisible particle, the basic unit of matter that takes part in chemical reactions.

The atom and electricity

The simple Daltonian view of the atom was overturned in 1897 when J. J. Thomson (1856–1940) discovered that atoms could emit even smaller particles of negative electricity (later called electrons) [5]. Clearly the atom itself must have some form of internal structure. Thomson's discovery also implied that an atom must also contain positive electricity. He suggested that electrons were like currants dispersed throughout a positively charged bun. This model failed to explain a number of the properties of atoms, but a better one had made use of the discovery of radioactivity by Antoine Becquerel (1852–1908). He found that certain heavy atoms spontaneously emit radiation. Three forms of this are now known: beta rays (negatively charged electrons), alpha particles (positively charged helium nuclei consisting of two protons and two neutrons) and gamma rays (short-wave X-rays).

The Rutherford model

In 1911 Ernest Rutherford (1871–1937) produced an entirely new model of the atom based on the results of his own experiments and those of Hans Geiger (1882–1945) and his co-workers (who measured the scatter of alpha particles when shot at gold foil). Rutherford's suggestion was that the positive charge and most of the mass of the atom were concentrated in a central nucleus and that the electrons revolve around it. We now know that the atom is mostly empty space with a minute central nucleus some tens of thousands of times smaller than the atom. The atoms themselves are extremely small – ten million of them side by side would form a line measuring only about 1mm (0.039in).

Rutherford later discovered that the positive charge of the nucleus is carried by particles 1,846 times heavier than electrons, he christened them protons. The charge of the proton is equal, but opposite, to that of the electron. A hydrogen atom consists of a single positively charged proton (the nucleus) with one electron travelling in an orbit around it [Key].

Heavier atoms have increasing numbers of protons in their nuclei, but the number of protons in the nucleus (called the atomic number) is always balanced by an equal number of orbiting electrons. It was later discovered that all atoms except hydrogen have another type of particle in their nuclei. These are uncharged particles (and are therefore called neutrons) and they have almost the same mass as the proton.

Quantum theory and spectroscopy

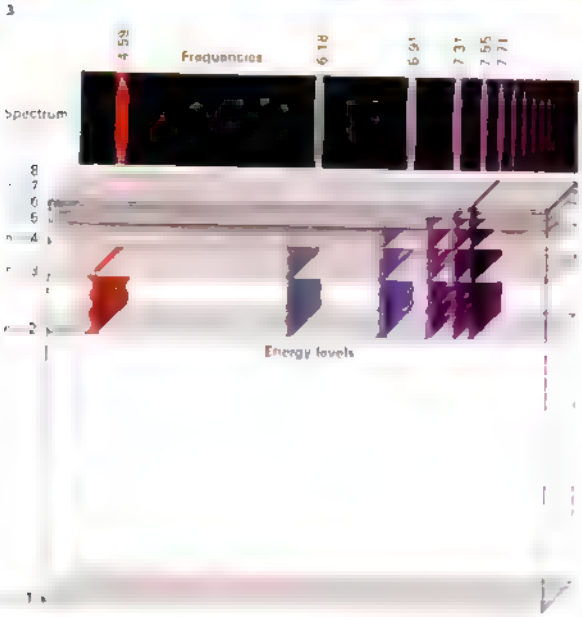
Two other fields of investigation helped the Danish physicist Niels Bohr (1885–1962) to construct the next important atomic model

CONNECTIONS

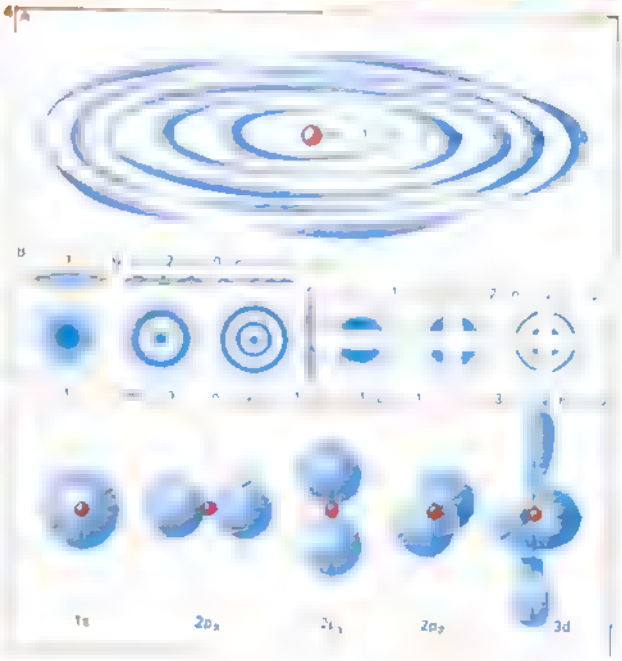
See also

1 The city of Miletus was the first known home of natural philosophy. Thales (c. 630 BC) was born there. He was a member of the Ionian School, the earliest known in Greek philosophy. He discovered the electrical properties of amber. Anaximander also dwelt there, as did Leucippus (c. 400 BC), credited by Aristotle with atomic theory that formed a central part in the evolution of Western scientific thought.

2 Modern science pictures matter as having a dual existence as waves and particles. Waves are seen on the sea and when a pebble drops into a pond, and sound and electromagnetic radiation such as light and X rays are known to travel as waves. The wave theory of atomic particles such as electrons, protons and neutrons has led to an improved understanding of atoms and nuclei.



3 Lines in the spectrum can be observed in light given out by incandescent elements. These are emission lines, which result from the emission of light by atoms. One of the successes of the Bohr theory was its ability to explain the wavelengths of the lines in the spectrum of hydrogen in terms of electron energy level changes.



4 Possible orbits of an electron around an atomic nucleus can be pictured (A) as circles that exactly accommodate a whole number of wave lengths, denoted by the principal quantum number n. A two-dimensional analogy (a vibrating drum skin) is described by two quantum numbers n and l (B), and the shape of a real atom (C) in terms of three (n, l and m)

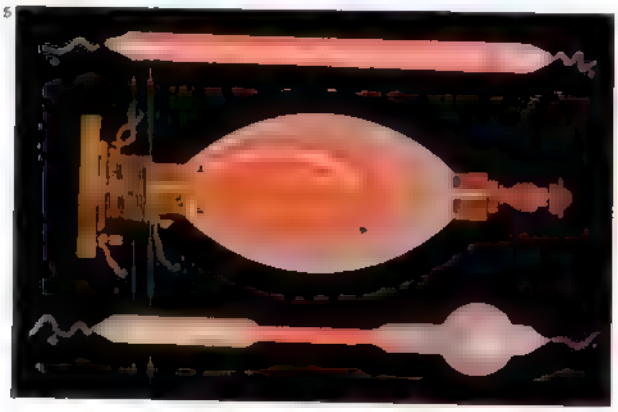
[Key]. The first was the quantum theory, the other was the science of spectroscopy. Quantum theory was proposed by Max Planck (1858–1947) [7] in 1900 as a way of explaining the emission of heat (and light) by a hot body. He realized that energy can be emitted and absorbed only in discontinuous amounts, discrete "packets" of energy that he called quanta.

Spectroscopy began when Isaac Newton (1642–1727) passed a ray of sunlight through a glass prism, breaking the ray into all the colours of the visible spectrum. In 1814 Joseph von Fraunhofer (1787–1826) had discovered that the spectrum of sunlight contains a number of black lines, which were later found to coincide with the position of coloured lines in the spectrum formed by electric discharge in hydrogen gas [3]. Bohr postulated that the circulating electron in an atom of hydrogen can exist only in fixed orbits [4A] and that the spectral lines correspond to the absorption (black lines) or emission (coloured lines) of a quantum of energy when this electron jumps from one fixed orbit to another. This theory, later modified by Arnold Sommerfeld (1868–1951), has been extremely successful in explaining the hydrogen spectrum.

Modern developments of the quantum theory suggest that the fixed orbits of Bohr should be visualized less precisely and that the position of an atomic electron should be treated as a probability that it will be in a certain place at a certain time. This treatment, known as quantum mechanics [2], was largely the work of Werner Heisenberg (1901–76) and Erwin Schrödinger (1887–1961). The substitution of a probability for a fixed orbit is a reflection of Heisenberg's uncertainty principle: if the momentum of a particle is known precisely there must be an uncertainty as to its position. In wave mechanics invented by Louis de Broglie (1892–), matter in the form of atomic particles is treated like light in that some of its properties are best explained in terms of particles, others in terms of waves. A stream of electrons behave like particles in cathode rays and like waves in an electron microscope. But for the purposes of chemistry the concept of the atom as the smallest unit of matter that can take part in reactions remains supreme.

The pictorial representation of the model of the atom proposed by Niels Bohr is an established part of the iconography of modern physics, even though Bohr's ideas have largely been superseded by various forms of quantum mechanics. This example is a hydrogen atom.

5



5 Geissler tubes, Victorian toys for adults, depended on electronic rays in a near-vacuum long before the principles of cathode rays were understood. The study of the rays, principally by J. J. Thomson, a British physicist, was the crucial one in the elucidation of the structure of the atom, by establishing the mass and charge of the electron, in conjunction with other experiments.

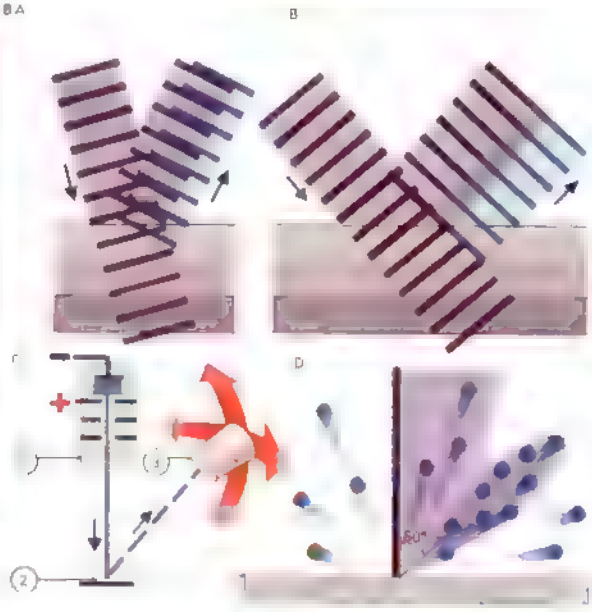
6

6 The new ideas about atomic physics were brought together at a series of conferences, such as this Solvay meeting at Brussels in 1911, attended by Bohr, Rutherford, Planck, Curie and others.

7

7 Max Planck suggested in 1900 that light was absorbed and emitted in packets or "quanta", with energies proportional to the frequency of light. This is known as the quantum theory.

8 A B



8 When waves are reflected from parallel surfaces they are out of step [A] or in step [B]. An electron beam [C] from a gun [1] can be reflected from nickel [2] into a detector [3] and the angles plotted [D].

9 Erwin Schrödinger played a principal part in the mathematical development of the modern model of the atom. He developed wave mechanics from de Broglie's picture of wave particle duality.

9

Nuclear physics

Nuclear energy plays a decisive part in shaping the modern world, nuclear weapons not only haunt the statesman but cast a threatening shadow on every living person [Key]. And while the mirage of limitless nuclear power attracts a civilization hungry for energy, the disposal of radioactive waste threatens lasting pollution of the world. In fact, life has always depended on nuclear energy: nuclear fusion heats the sun [1], and radioactivity in the earth [2] heats the liquid core and contributes to the mobility of the continental plates. Nuclear energy is derived from splitting the atomic nucleus by radioactivity or fission, and secondly from fusion, which is the joining of a pair of light nuclei.

Radioactivity: its discovery and source

Radioactivity was discovered by Antoine Becquerel (1852–1908). With the isolation of radium it became clear that enormous amounts of energy were involved. Radium decays over many decades and in fact contains 2×10^5 times as much energy as an equal mass of coal. A nucleus, a few times 10^{-12}cm in diameter, is made of protons (positively charged particles) and neutrons (neutral particles of nearly equal mass to the proton). Hydrogen is unique in having a single proton (and no neutrons) in its nucleus. Most elements consist of a mixture of isotopes, whose nuclei differ in their numbers of neutrons. The total number of constituents (protons and neutrons) in a particular isotope is indicated by a superscript, for example He^4. It is on the properties of individual isotopes that nuclear power depends.

Deliberate transformation of one nucleus into another was achieved by Ernest Rutherford (1871–1937) [4] in 1919: $He^4 + N^{14} \rightarrow O^{17} + H^1$. In words, an alpha-particle (the nucleus of helium) and a nitrogen nucleus momentarily combine and then split into the oxygen isotope O^{17} and a proton.

As mass spectrometers [7] – instruments that measure the individual masses of ions and thus of the nuclei – became more accurate, it was found that the masses of the nuclei of the various isotopes were not equal to the sum of the masses of the constituent protons and neutrons. This discrepancy, according to Einstein's relativity formula, $E = mc^2$, is the source of nuclear energy. Modern theory views the nucleus to be rather like a liquid droplet of neutrons and protons. Any such system tends to decay – that is to say transfers itself – into a state of lower energy. If it does so by breaking into two nearly equal parts, that is called fission, if a nuclei gives off one or more particles, that is radioactivity, if two nuclei join together, that is fusion. Two nuclei are both positively charged so one of them must be accelerated to high speed to achieve fusion, or both must be moving fast due to high temperature.

Generating nuclear energy

The generation of nuclear energy in large quantities by fission requires a chain reaction, first achieved with uranium [6]. When a neutron is absorbed by the isotope uranium-235 it induces fission into two major fragments and two or three neutrons. If, on average, more than one neutron from each fission causes a second fission, the process may accelerate exponentially into a runaway chain reaction.

In order to generate electricity the pro-

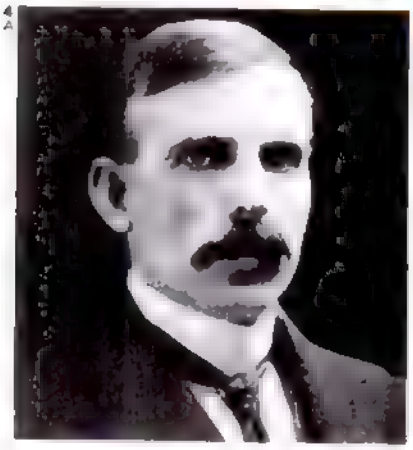
CONNECTIONS

See also

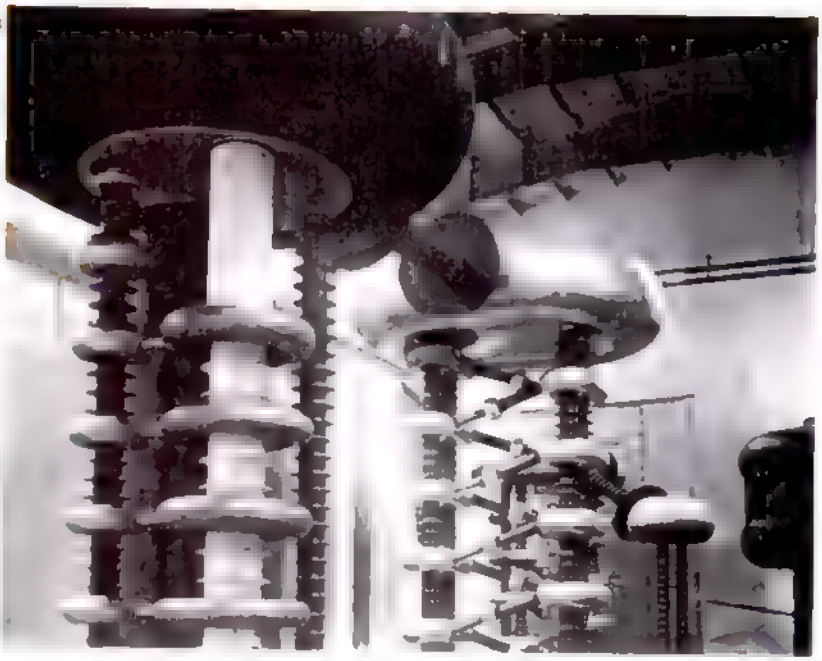
1 **The sun** is powered by nuclear fusion which needs temperatures of the order of hundreds of millions of degrees centigrade. On earth the necessary conditions and temperatures for fusion reactions have so far been achieved only in bombs. Nuclear energy by controlled fusion remains a dream.

2 **Volcanic energy** is provided in part by radioactivity that is naturally present. This is such a powerful source of energy that radioactive atoms slowly decaying inside the earth, contribute significantly to maintaining the temperature of the molten core. Thus natural radioactivity indirectly helps power volcanic eruptions.

3 **Marie Curie** (1867–1934) and her husband Pierre Curie (1859–1906) formed one of the most famous husband and wife teams in the history of science. During investigations of the radiations given off by uranium, they found an inexplicably high level of radiation. Through painstaking chemical detective work they tracked down and isolated its source, the radioactive elements radium and polonium.



4 **Ernest Rutherford** (A), a New Zealand-born physicist who first worked at Cambridge University in 1903, established the nuclear theory of the atom in 1911 and later achieved the first splitting of the atom when he produced protons from the nuclei of nitrogen atoms. He and his team used remarkably simple (including "home-made") apparatus at the Cavendish Laboratory [8] and were able to change our picture of the structure of atoms.



cess must be slowed down and controlled and means must be provided for removing the heat. An atomic pile to produce electric power is merely a special kind of furnace. There is a higher probability of fission occurring in U-235 if the neutron absorbed is moving relatively slowly, about 2km (1.2 miles) per second. For this reason special materials, called moderators, are incorporated in the atomic pile to slow the neutrons.

Atomic piles may be classified according to the material used as moderator, typically graphite, water or heavy water. Heavy water is water in which the hydrogen is a heavy isotope, deuterium, and it absorbs far fewer neutrons than ordinary water.

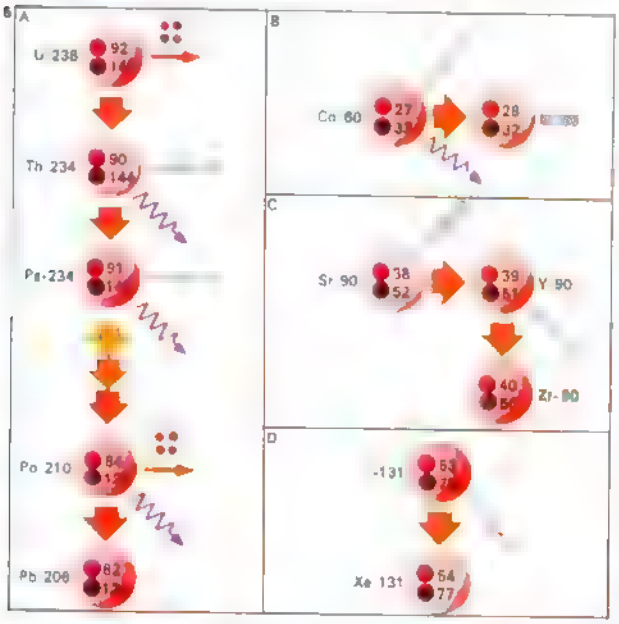
The amount of fissile material brought together is crucial for a sustained chain reaction. If more neutrons are lost by absorption or escape than are produced, the reaction will not be self-sustaining. If more neutrons are produced than are lost on average, a self-sustaining and expanding reaction occurs. The smallest amount in which fission is self-sustaining is called the critical mass. In an atomic pile it is necessary to keep the flux of neutrons nearly in balance and constant. To control a pile, rods of neutron-absorbing material can be moved in and out [8].

Fast reactors and their make-up

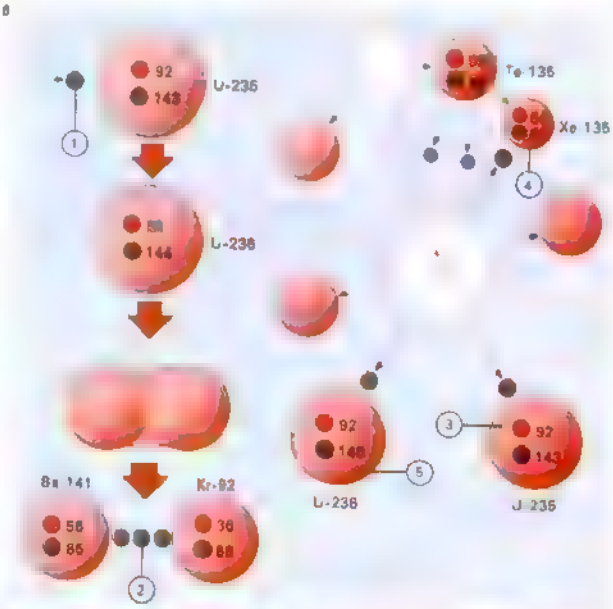
Structural supports in atomic piles are made of materials that absorb as few neutrons as possible. Fast reactors have a small core of fissile material and no moderator to slow the neutrons. There is little absorbent material and few neutrons are wasted. Natural uranium is 99.3 per cent U-238 and 0.7 per cent U-235. While U-235 is spontaneously fissionable, U-238 is not, but after absorbing a neutron can decay radioactively to plutonium Pu-239, which is. Atomic bombs and fast reactors require fairly pure fissionable material so U-235 must be separated out – for example in a giant diffusion plant – or U-238 turned into Pu-239 in a reactor and separated chemically. In a fast reactor, the fissile core is surrounded by a blanket of natural uranium so that neutrons escaping from the core can turn U-238 into Pu-239. If more fissionable material is made than consumed, the reactor, is a breeder reactor.

KEY

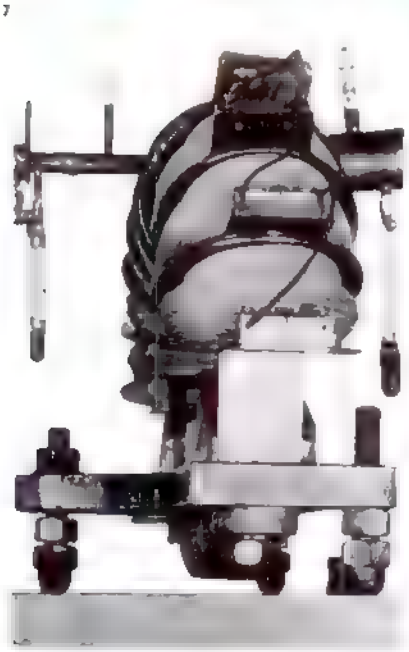
The mushroom cloud of an atomic explosion haunts our civilization. Although the spread of nuclear weapons has been banned by treaties, and some nations have accepted technical limitations on testing, the number of nations with access to nuclear weaponry continues to grow.



5 The nucleus of an atom, with protons (red) and neutrons (brown), may change, giving radioactivity gamma rays (electromagnetic radiation, violet), electrons or beta rays (grey), positrons, or alpha particles (helium ions, orange). Naturally radioactive uranium 238 [A] decays as shown to form lead [B, shows the decay of cobalt-60, [C] strontium 90 and [D] iodine-131.



6 When uranium-235 is hit by a slow neutron [1] it may split by stages and release energy and more neutrons [2]. One of these may strike more uranium-235 [3] and lead to a chain reaction, or be absorbed by other atoms [4] or U-238 [5].



7 The mass spectrograph of Francis Aston (1877–1945) showed that elements are formed of separate isotopes, each nearly an integral multiple of the mass of a proton. Later spectrographs gave exact measurements of nuclear masses and are used to distinguish isotopes.

8 Nuclear reactors are the power houses of the future and to some extent of the present. But the formidable problems they create in disposal of radioactive wastes have not yet been satisfactorily solved. Here engineers are carefully stacking the rods that go to make up the central core of an atomic pile. Control rods are also inserted to regulate the reaction rate.

Floating insects such as water boatmen demonstrate an interesting use of surface tension. Any one atom within the body of the water has equal and opposite forces exerted on either side of it in any direction. But at the surface there is a net force pulling the atoms into the body of the water because the atmosphere above the surface is less dense than the water. As a result, the surface appears to behave like an elastic skin pulled across the water. This surface tension effect can be used by the insect to keep itself afloat.

Even fairground thrills

Beyond the atom

One of the characteristic features of science is the way in which it attempts to explain a collection of different phenomena in terms of a few basic concepts. A striking example is the atomic theory of John Dalton (1766–1844) in which many different substances are considered to be made up of a few different types of atom. According to this view atoms are the fundamental "building blocks" of all matter.

In the late nineteenth and early twentieth centuries evidence accumulated to show that atoms themselves have an internal structure. By 1932 it had been realized that atoms are combinations of sub-atomic particles: protons and neutrons (together forming a small positively charged nucleus) with orbiting negatively charged electrons.

Interactions between particles

To give a full description of matter it is necessary to describe not only the particles but also the way in which they are held together – that is, the way in which they interact with one another. Four types of interaction are recognized; two of these are fairly well known because they are observed in matter in bulk as well as on the atomic scale. The gravitational interaction [1] produces an attraction between objects that depends on their masses. It is an extremely weak effect and plays no part in the binding within atoms, but it is responsible for the forces between heavenly bodies. The electromagnetic interaction [2] occurs between particles that have an electric charge. This force is many millions of times stronger than the gravitational effect and is responsible for the force of attraction between the nuclei of atoms and the orbiting electrons.

Within the nucleus itself a quite different effect must occur. Here neutrons and protons are held together strongly in spite of the electromagnetic repulsion between them. This strong interaction is independent of charge for it acts between neutrons as well as protons, and is about 7,000 times stronger than the electromagnetic interaction. Moreover, it falls off sharply with distance – its influence extends only over distances comparable with the dimensions of the atomic nucleus, generally less than 10^{-13} cm.

The fourth type, known as the weak interaction, is about one thousandth of the strength of the electromagnetic interaction. It is observed in certain processes in which transformations of particles occur, as in beta decay, where a neutron changes into a proton, an electron, and an antineutrino.

Fields of force

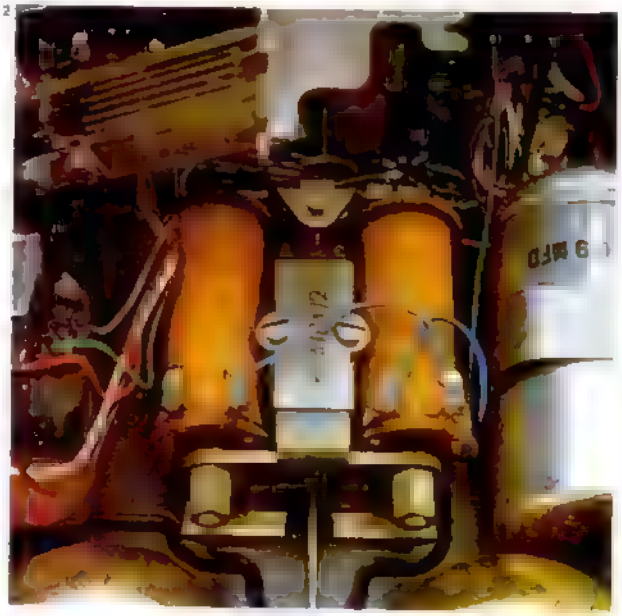
The four types of interaction take place through free space. One way of explaining this action-at-a-distance uses the idea of a field of force. A charged particle, for instance, is thought of as affecting the surrounding space in such a way that another charged particle placed in this region experiences a force. The region of influence is called an electromagnetic field. Similarly a mass has an associated gravitational field in the space surrounding it.

A different model, based on quantum mechanics, uses the idea of an exchange of virtual particles. Two charged particles interact by emitting and absorbing photons (particles of light). Gravitational interaction is similarly explained by exchange of

CONNECTIONS

See also

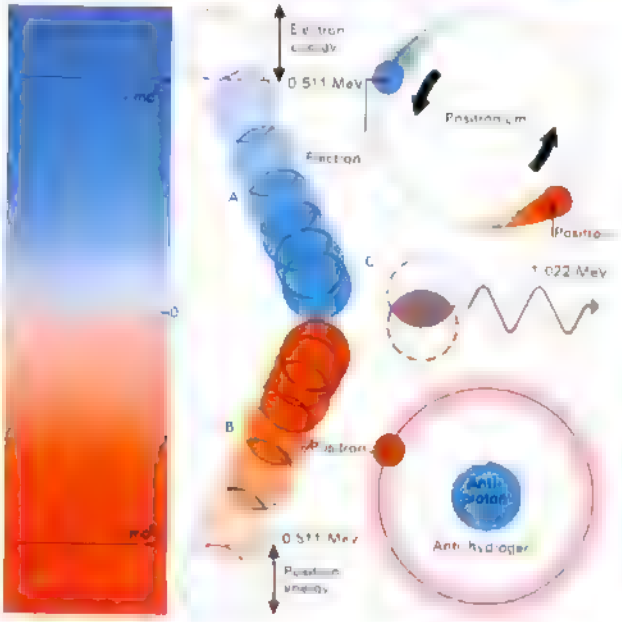
1 The force of gravity is encountered in all its power and immediacy by a weightlifter. The gravitational force was the first to be studied quantitatively and the first to receive in Isaac Newton's *Principia* (1687), detailed discussion of its theoretical principles.



2 Electromagnetism (used in an electric bell) was the second interaction of which man became aware. Magnetism and static electricity were known earlier but the combination of electricity and magnetism in a single theory was achieved only in the nineteenth century by the work of Oersted, Faraday and Maxwell.

3 Hideki Yukawa (1907–) predicted a particle, later identified as the pi meson, as the quantum of the strong nuclear force. It was found in cosmic ray photographs by C. Powell (1903–69) at Bristol University.

4 Enrico Fermi (1901–54) developed the theory of beta decay, which depends on weak interaction. He was largely responsible for the development of the first atomic pile (nuclear reactor).



5 The existence of a particle equal in mass to an electron [A] but having negative energy was predicted by Paul Dirac (1902–). The discovery of the positron [B] confirmed this. If the two meet, they annihilate each other, releasing their combined mass-energy [C]. Positronium – an electron and positron in orbit – is known to exist briefly and as other anti-particles are known it is possible to imagine anti-atoms such as anti-hydrogen in some other world.

hypothetical particles called gravitons. In 1935 Hideki Yukawa [3] suggested that the strong interaction holding the nucleus together was due to the exchange of a particle with a mass between that of the electron and the proton. This particle is now known as the pi meson (or pion). Another particle, the intermediate vector boson, has been suggested as being responsible for weak interactions, but so far scientists have been unable to prove its existence.

Other fundamental particles

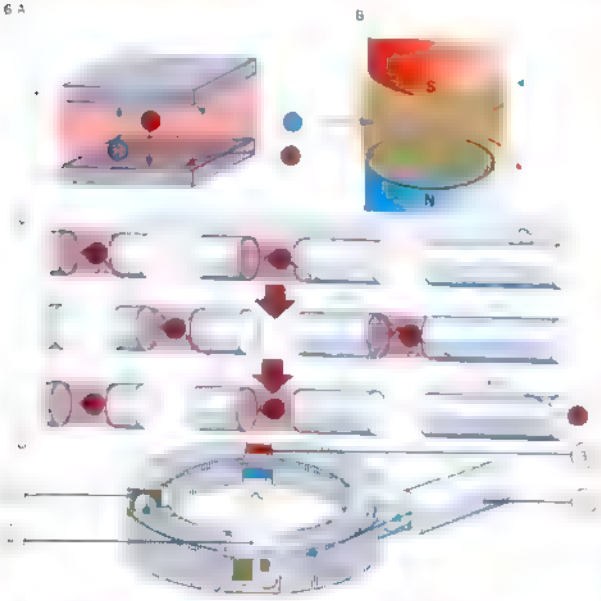
In 1932 only three particles were necessary to explain atomic structure. Since then the situation has been complicated by the discovery of many more particles through work on cosmic rays and experiments using particle accelerators [6, 8]. It is found that high energy collisions between particles lead to the production of new ones. Now more than 200 are known, most of them very unstable [7]. They are characterized by their mass and charge. They also have other characteristic properties, such as average lifetime, that describe the ways in which they interact.

The numerous sub-atomic particles are classified into groups: particles that partake in strong interactions are called hadrons (including nucleons, hyperons, and mesons). Particles that do not take part in strong interactions are called leptons (including electrons and neutrinos).

The problem of high-energy physics is to produce a single theory explaining the existence and behaviour of this multitude of particles. One suggestion is that the particles themselves are made up of even more basic particles. It is possible, for instance, to describe all nucleons as combinations of three particles called quarks. These have charges that are one-third or two-thirds the size of the electron charge. According to this theory protons and neutrons, for example, are not completely indivisible.

Another goal is the creation of a single theory to account for all the types of interaction. So far some success has been achieved in unifying the electromagnetic and weak interactions, but a single mathematical theory encompassing all four types of interactions is still far off.

The essence of matter summed up in the Chinese symbol of yin and yang was a symmetry of complementary principles – aptly representing the modern theory of particle-wave duality.

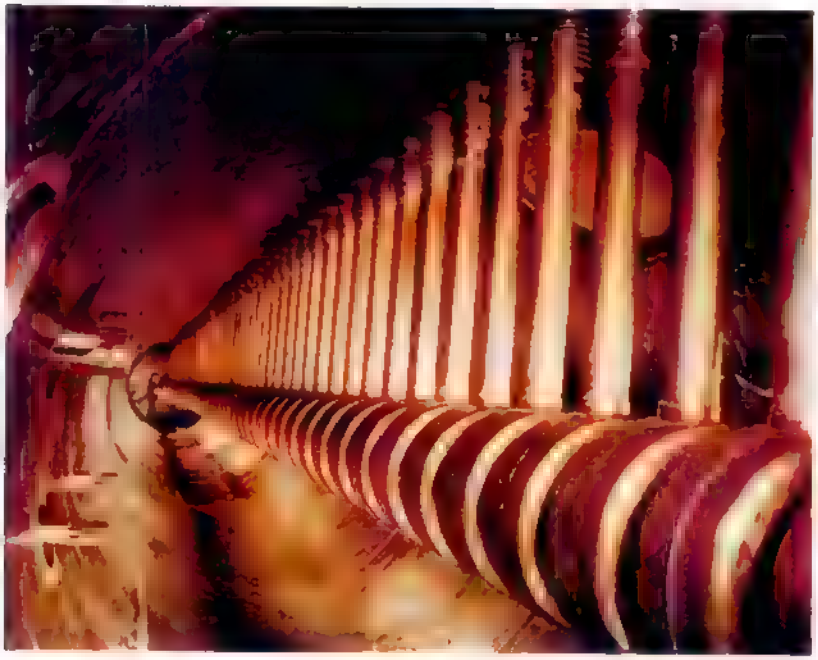


6 Particle accelerators use the principle that an electric field [A] accelerates or deflects positive (red) or negative (blue) particles parallel with the direction of the field, whereas a magnetic field [B] makes them curve at right angles to the field. In a "drift-tube accelerator" [C], oscillating electric fields speed up at the same rate as the particles. A synchrotron [D] is used for particles brought close to the speed of light by a linear accelerator [1]. A magnetic field [2], increasing to balance the growth of centrifugal force, causes particles to circulate through an accelerating field [3], seen through a viewer [4].

7

Type of particle	Symbol	Mass in units of electron mass	Electric charge
Photon	γ	0	0
Leptons			
Neutrino (electronic)	ν_e	0	0
Antineutrino (electronic)	$\bar{\nu}_e$	0	0
Neutrino (mesonic)	ν_μ	0	0
Antineutrino (mesonic)	$\bar{\nu}_\mu$	0	0
Electron	e^-	1	−1
Positron	e^+	1	+1
Mu (muon) plus	μ^+	207	+1
Mu (muon)	μ^-	207	−1
Mesons			
Pi plus	π^+	273	+1
Pi minus	π^-	273	−1
Pi zero	π^0	264	0
K plus	K^+	967	+1
K minus	K^-	967	−1
K zero	K^0	974	0
Anti-K zero	$\bar{K}^0$	974	0
Nucleons			
Proton	p^+	1 836	+1
Antiproton	$\bar{p}$	1 836	−1
Neutron	n	1 839	0
Antineutron	$\bar{n}$	1,839	0
Hyperons			
Lambda zero	Λ^0	2 183	0
Antilambda zero	$\bar{\Lambda}^0$	2 183	0
Sigma zero	Σ^0	2 332	0
Antisigma zero	$\bar{\Sigma}^0$	2,332	0
Sigma minus	Σ^-	2 328	−1
Antisigma minus	$\bar{\Sigma}^-$	2 341	+1
Sigma plus	Σ^+	2 328	+1
Antisigma plus	$\bar{\Sigma}^+$	2 328	−1
Xi plus	Ξ^+	2 566	0
Antixi zero	$\bar{\Xi}^0$	2 580	0
Xi minus	Ξ^-	2,580	−1
Antixi minus	$\bar{\Xi}^-$	2 582	+1

7 Fundamental particles are now so numerous that the term fundamental seems inappropriate but there is insufficient information for a unified theory and some vital clues may be missing. Perhaps one day order will be brought to the apparent chaos of information, just as Niels Bohr clarified atomic theory.



8 The ring cyclotron was invented by Professor Ernest Lawrence (1901–58) at the University of California in 1930. Its vacuum chamber, in which charged particles were accelerated by being circulated past two D-shaped electrodes with a high-frequency voltage, measured only 9.8 cm (4in) across. The giant machine now at CERN in Geneva has a diameter of 4.8km (3 miles). Accelerated to high speeds in cyclotrons, particles can be used to bombard nuclei – with the creation of other particles.

The nature of energy

Energy is required for work to be done – work being the operation of a force over a distance. Thus energy is expended when a golf ball is struck, when a dumbbell is lifted, when a spring is compressed or stretched, when a bomb explodes and when electrons flow in a wire as an electric current.

Additionally, energy is needed to raise the temperature of any substance and living organisms need energy for movement and growth. Green plants get their energy as light from the sun [Key], which they utilize by photosynthesis. Animals use the chemical energy of food – the energy of plants or other animals that they eat. From these examples it can be seen that energy exists in many forms. To a scientist, it exists in only two distinct ways: as potential (or stored) energy, and as kinetic energy (which a body possesses by virtue of its motion).

The essence of potential energy

Potential energy can be considered as stored energy. The potential energy of food and fuels such as coal and oil, for example, is the chemical energy stored in these materials. The potential energy in the water of a dam is equivalent to stored gravitational energy, the earth tries to "pull down" the raised mass of water with a force proportional to its mass [1] the force of gravity.

In a coiled spring, the potential energy is proportional to the square of the compression or extension. A copper sphere insulated from electrical leakage can be charged with static electricity (unmoving electrons) and the electrical potential energy of the sphere is determined by the amount of static electric charge and the associated voltage.

What is internal energy?

The total energy of any system is called its internal energy. This quantity is not usually measurable and not all of it can be used to do work. A hot object does work in cooling, for example, but even if it is cooled to near absolute zero, namely −273°C (−460°F), its molecules still possess most of their internal energy. The potential energy of a sytem is measured not as the total internal energy but as the part of it available to do work.

When matter is in motion it is said to possess kinetic energy. For this reason, molecules of a gas always have kinetic energy because they are always moving. The temperature of a gas is a measure of the average kinetic energy of its moving molecules – the faster they move, the higher the temperature.

The pressure of a gas is also a measure of its kinetic energy, because the pressure is a measure of the number and energy of the collisions made by gas molecules on the walls of its container. Gas pressure is often used, as in pneumatic drills and lifts, to do useful work. Finally, the kinetic energy of a gas molecule, or of any other moving object, can be expressed in absolute mathematical terms as $\frac{1}{2}mv^2$, where m is the mass of the object and v is its velocity.

Conservation of energy

Heat energy can be regarded as energy of movement. To take the example already used, a gas has kinetic energy proportional to its absolute temperature (its temperature above absolute zero). If the same gas heats an object, the molecules of the object gain

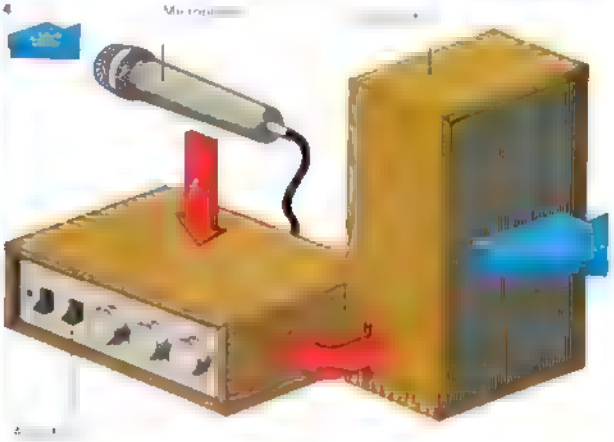
CONNECTIONS

See also

1 A dam holds water at a height and when the water is released its potential energy changes to kinetic energy that can be converted into useful electrical energy by water turbine generators. Other hydroelectric processes rely on the kinetic energy of water moving in the form of waves and tides.

2 In batteries the energy of chemical reactions is converted into electrical energy, which is used for many familiar purposes. Batteries run down as their chemical activity declines. Accumulator batteries (secondary cells) can be recharged by electricity, which is converted back to stored chemical energy.

3 A nuclear power station uses the great energy released when the nucleus of an atom disintegrates. Under controlled conditions the nuclear energy of uranium or plutonium is carefully released, mostly as heat, which is used to raise the temperature of water in boilers and produce steam. This heat energy is converted in steam turbines into mechanical energy, then to electrical energy in generators.



4 In a microphone-loudspeaker system a double energy conversion takes place. Sound energy is converted by a microphone into electrical energy. The loudspeaker reverses the conversion process.

kinetic energy. The stirrer [illegible] kinetic energy takes the [illegible] of heat from the hot [illegible] the [illegible] body.

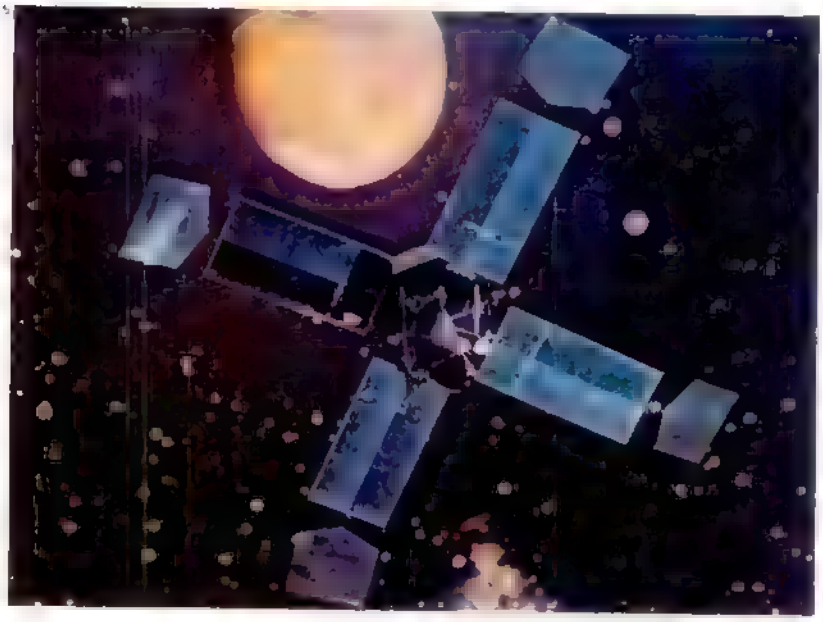
[illegible] energy is usually thought of in such terms as the heat of burning coal, the electricity flowing in electric fires and the mechanical energy produced by burning petrol in a car engine. But in each of these familiar examples energy has had to be converted from one form to another to be of practical use. The chemical potential energy [illegible] used by burning the coal and is thereby converted into useful hot gases and radiant energy; in a power station these are further converted into useful electrical energy by a system of water boilers, steam turbines and electricity generators. The chemical potential energy of motor fuel, is released by rapid burning as kinetic energy of gas, which is translated into the useful mechanical energy that propels the car.

Energy conversions always involve a loss; no conversion is 100 per cent efficient. A coal fire, for example, releases only about 20 per cent of the chemical energy of the coal as useful heat. An electric motor converts about 80 per cent of the electrical energy supplied to it into mechanical energy [7].

The principle of conservation of energy is usually stated as "energy can neither be created nor destroyed". In energy conversions the amount of energy output to do useful work is always less than the input energy. The total energy of the system, however, always remains the same, the "missing" energy being wasted energy. Not all the electricity flowing through a lamp filament, for instance, is converted into light, most being wastefully converted to heat. The heat and light together are equivalent to the input electricity and so energy is conserved. But after the conversion, less energy is available to do work, as a whole, the useful energy is at a lower level.

What happens in this particular instance is true for the sum of all energy reactions in the universe at any one time; the overall result of these reactions being a degradation in energy level. Perhaps at some remote time in the future, all energy will have degraded to a level where no work can be done, the universe will then have "run down".

The sun is the ultimate source of energy for all life on earth. Here it is reappearing from behind the moon after an eclipse.

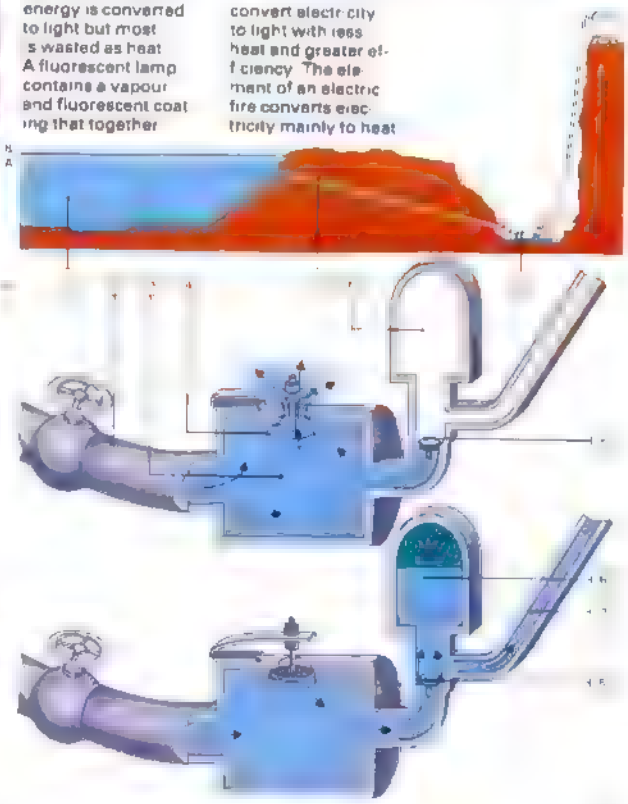


5 Solar cells, as fitted on this satellite, convert radiant energy from the sun directly into electrical energy. A solar cell contains a wafer of semiconductor material usually silicon. This is made in such a way that when light falls upon it, electrons move in one direction within it and 'holes' (positively charged regions) in the other direction. In a circuit each solar cell produces about half a volt.

6 In the filament of an electric light bulb some electrical energy is converted to light but most is wasted as heat. A fluorescent lamp contains a vapour and fluorescent coating that together convert electricity to light with less heat and greater efficiency. The element of an electric fire converts electricity mainly to heat.

7 An electric motor converts electrical energy into mechanical or kinetic energy. Most familiar is the rotary type in which a rotor revolves within a fixed stator.

8 The hydraulic ram raises water by converting its kinetic energy into gravitational potential energy. A lake or reservoir [A] supplies water [1] through a pipe [2] to the ram chamber, which is at a lower level ensuring that the water has adequate kinetic energy on reaching the chamber [3]. This fills with water [B] which briefly escapes through a spring valve [4] before closing it. The water passes on round a one-way valve [5] into a second chamber [6] where air is first compressed by the water [C] then re-expands to force water up the delivery pipe [7]. A back surge allows the spring valve to reopen and the process to be repeated.



Statics and forces

The sudden movement of an object with seemingly no cause – for example, the unexpected movement of a table in the middle of a room – would obviously cause consternation to the observer. Most people would probably think a trick was responsible for such a happening because they expect a cause for this effect of motion. The scientific name for the cause is "force" – it is anything that causes an object to start to move when it has been at rest, or vice versa. A force is also required to change a condition of motion that already exists – to change the direction of an already existing motion or to alter the velocity of the motion. Once made to move, an object continues to move without stopping or changing direction until it is acted on by another force. This idea has now become almost self-evident following its original expression and generalization by Isaac Newton in the late 1600s, and is generally known in physics as Newton's first law of motion.

In many cases the cause that stops, or modifies, already existing motion is provided by the force of friction. This acts in the direction opposite to the movement of an object. It is produced by the rubbing together of the surface of the moving object and the surface it is moving on or, in a gas or a liquid, the medium it is moving through.

States of equilibrium

When several forces act at the same time on the same object, each tries to move the object along a line pointing in its own direction at a rate that depends on the size of the applied force. If it happens that the object does not move as a result of all these forces, then it is said to be in equilibrium [5].

The magnitude and direction of any one of the forces is balanced by the total effect of the other forces and there is no resultant movement. The study of forces applied to objects in a state of equilibrium is called statics (as opposed to dynamics, the study of forces acting on moving objects).

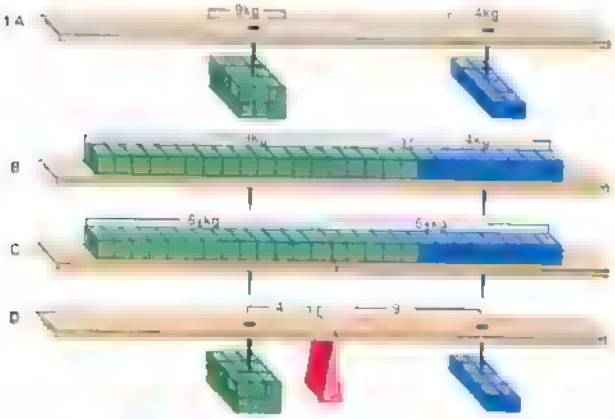
Someone sitting still on a chair is an everyday example of an object in equilibrium – the upward force of the chair on the person balances the downward force of the earth's gravitational attraction that is trying to pull the person through the chair to the floor. A tree standing upright in the ground is a similar example – its downward gravitational force, commonly known as its weight, is balanced by the upward force of the ground in which it is rooted. As a consequence of this equilibrium, neither the person nor the tree is in motion up or down. This state of affairs can be modified only if another external force is brought into the arrangement – for example, by the person moving about on the chair and possibly making it topple or, in a most extreme case, by chopping down the tree [7]. Thousands of everyday objects stay where they are and do not move unaided because of the existence of this equilibrium of forces both in magnitude and direction.

Moments and levers

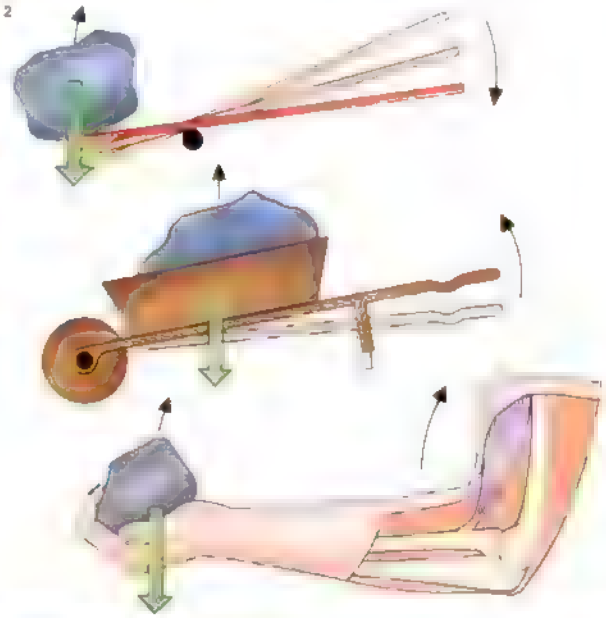
So far, only those forces that try to move objects along straight-line paths have been considered. There are many other forces, however, that can act on objects and try to rotate them around a central point. These forces have an effectiveness that depends on how far from the central point they are acting. Everyone knows that a much greater

CONNECTIONS

See also

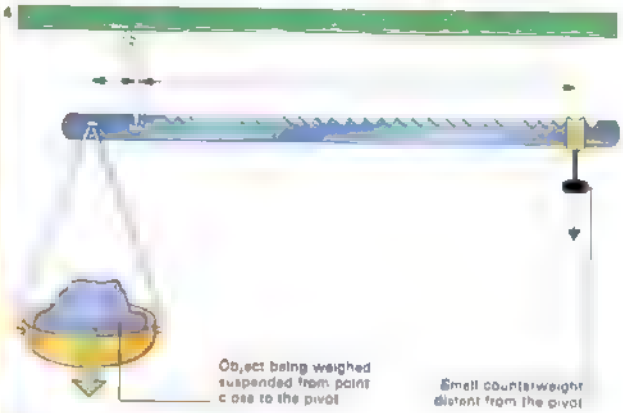


1 The balance point of the lever (itself assumed weightless) [A] can be found by dividing each load into equal small weights and distributing them along the lever [B], keeping the centre of gravity of each load where it was. The lever balances at the centre of gravity of the whole line. This point is halfway along the line [C] of equal weights at the distance from the weights that is inverse to the ratio of the weights themselves [D]. The product of the weight and its distance from the fulcrum (pivot) is the same on each side – the moments are equal and opposite.



2 The see-saw principle is used in the first order of levers and has the fulcrum, or balance-point, between the load and the applied effort. The wheelbarrow is a familiar example of the second order of levers, with the load between the fulcrum and the effort. Finally, the third order of levers, with the effort between the fulcrum and the load, is used when lifting a weight with the forearm – the elbow is the fulcrum. If both lever and load remain the same in all three cases, the principle of moments shows the required force is the greatest for the third order and least for the second order, which is the best form of leverage.

3 The force needed to turn this threshing machine is lessened if the animal pulls on the outer harness chain, because the leverage (that is force × perpendicular distance to the axis) is greater. But in doing so the animal must walk further.



4 A steelyard is a balance often used by butchers for weighing carcasses that are too heavy for ordinary scales. The small sliding jockey weight is moved along the long arm of the steelyard until the whole beam is horizontal. The sliding weight, at a greater distance from the fulcrum than the heavy load, can be much lighter. The principle involved here is the same as the first order of levers – like the see-saw. Its first recorded use was in 315 BC.

turning effect is achieved by pulling on a wrench or spanner that has a long, rather than a short, handle The combined effect of the magnitude of the force multiplied by the perpendicular distance from the turning point (called the axis of rotation) is called the moment of the force. The greater the applied force or the distance of action, the greater is the resulting moment or turning effect Simple machines that employ the phenomenon of moment of a force belong to one of the various classes of levers [2]

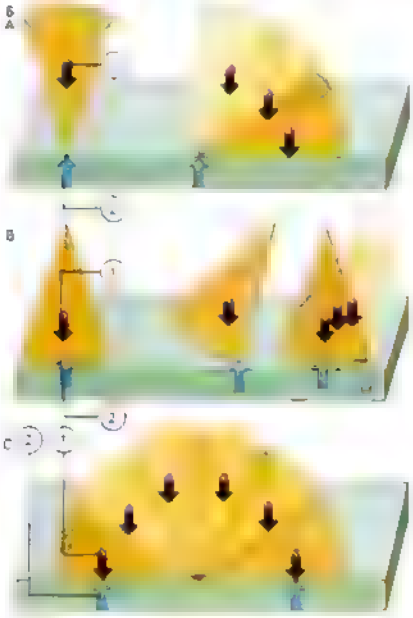
Of course several forces may simultaneously act on an object through more than one point within its boundary. In this case, each has its own moment about a particular axis, through which the object can still be in equilibrium if these moments produce no collective rotational effect. That is, if the total magnitude of the clockwise moments about the axis exactly balances that of the anti-clockwise moments, there is no movement This result is called the principle of moments and it may apply to an object at the same time as the equilibrium, mentioned previously, that arises from forces acting through a particular point trying to move the object in a straight line

For the example of a person sitting still on a chair and being in equilibrium for vertical motion, it is clear that if someone tries to tilt the chair backwards, the seated person will topple over unless someone or something pulls the chair in the opposite direction with an equivalent force. The vertical equilibrium through the point of contact of the chair with the floor acts at the same time as the equilibrium for the moments of the forces trying to twist the chair around at the same point

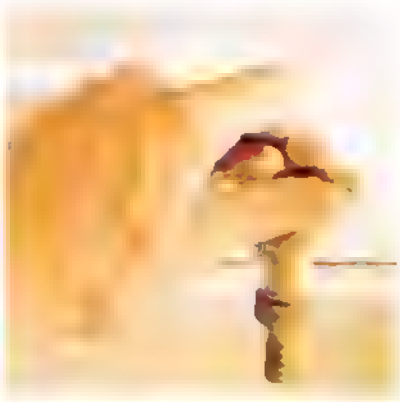
How couples operate

One further type of force that a study of statics includes is called a couple [6]. Actually this is the application of two equal forces arranged so that they both tend to rotate the object in the same direction. The couple produces only rotation with an *equal* moment about any point between the forces, and does not produce any straight-line motion Consequently it can be balanced only by another equal or opposite couple if an equilibrium state is to be achieved

The principle of moments s used in most of the activi ties in a children's playground The see-saw obvious y demonstrates the lever as shown here in addition, the um brella, roundabout and swings all use the moment of a force about their axis or fulcrum to achieve movement

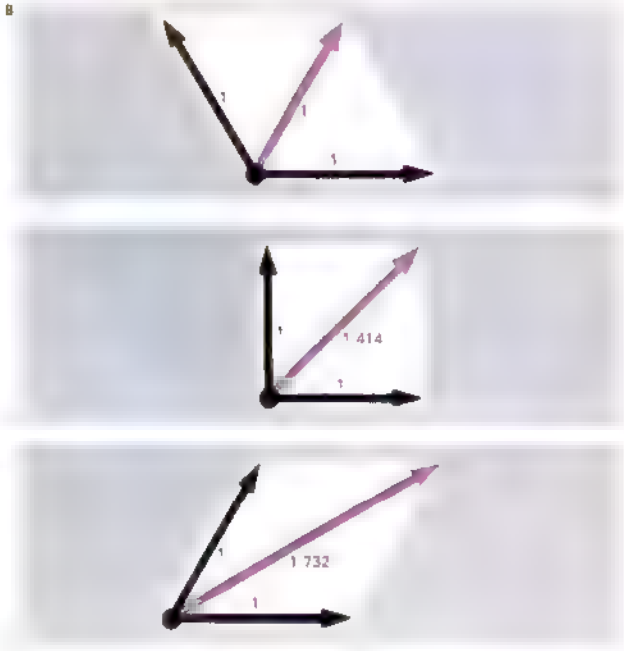


5 If a cone standing on its point [A] is sl ghtly d splaced, its weight produces a moment that continues to topple the cone about the point of contact This is an unstable equ librium position With the cone on its base [B], displacement produces a moment that restores the original position – a stable equilibrium If the cone lies on its side [C], displacement produces no moment since the weight and its reaction [1, 2] still act along the same l ne – the cone remains in its new position and is in neutral equi ibrium. The very low centre of gravity of the bus [D] keeps it in stable equi ibr um even for large displacements



6 The moment of a couple is equal to the product of one of the forces and the distance between the two. Th s hand is applying an equal force to each end of the top of a tap, producing a rotating effect about the cen tral axis to turn it on and off Both these forces are trying to turn the tap in the same d rection, and together form a couple, the larger the tap-top, the larger the couple

7 A tree stands in equilibrium (its centre of gravity acts downwards in a straight line through its base) unt l a wedge is cut into its trunk This starts to destroy the equilibr um by allowing the tree's weight to develop a toppling moment that is not balanced by an equal and opposite reaction.



8 Two forces act simultaneously on an object in different directions. The resultant force, and the direction of any subsequent movement s def ned by the diagonal of a parallelogram whose s des are drawn parallel to the applied forces with lengths pro port onal to their magnitudes. This is an appl cation of vector diagrams

Attraction and repulsion

To many people, the most mysterious natural forces are those that produce an effect on objects at great distances, reaching across even empty space without any material contact between the body producing the force and that being affected by it. This phenomenon is often called "action at a distance", and there are several fundamental forces of nature that act in this way. They are gravitational, magnetic and electric. Other fundamental forces act within the limits of the atomic nucleus, over short ranges only.

The inverse-square law

The three forces acting at a distance obey a common law that describes how the magnitude of the force depends on the object producing it, the object affected by it and the distance separating the two. It is called the inverse-square law [Key]. If the distance between two bodies is doubled, then the force between them falls by one-quarter (the inverse square), and so on.

Isaac Newton (1642–1727) was first to realize this type of law applied to the gravitational force of attraction when performing calculations on the speed of the moon in orbit round the earth. His law of gravity [1] states that "the attractive force between two bodies is proportional to the product of their masses divided by the square of the distance between them" and it should be noted that this force is a mutual one – there are equal and opposite forces on the two objects.

About 100 years later, in 1789, Henry Cavendish (1731–1810) used this theory of gravitation to produce the first estimate of the mass of the earth [2]. He was also responsible for one of the best experimental verifications (in 1785) of the use of the inverse-square law to describe the force between electrostatic charges (that is, electric charges at rest). Its use was further verified in 1870 by James Clerk Maxwell (1831–79) who showed that the inverse-square law was true to one part in 20,000. More modern methods have taken this limit to one in 1,000 million.

The volume of space within which the force exerted by a body produces a detectable effect (usually by causing movement of a second body) is normally called the field of force. The directions along which movement can occur are known as the lines of force. These are merely imaginary "lines" in space that are used to help describe the possible directions of any motion within the field – they spread out from the force-producing object, filling the complete field.

Gravitational and magnetic force

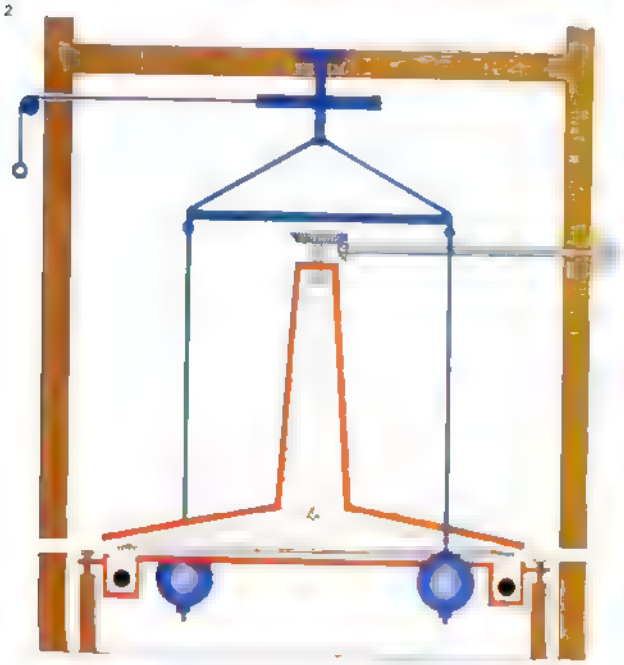
Any object with mass, however small, can produce a gravitational force field. The force is always positive because no object can have a negative mass. This indicates that the force of gravity can only be one of attraction. (A negative value for a force indicates repulsion, not attraction.)

It is gravitational attraction that gives all objects their weight, trying to pull them towards the centre of the earth [3], and which keeps the planets in orbit around the sun. In some circumstances, gravitational attraction can be balanced by an equal and opposite force to achieve a weightless condition. This is most commonly accomplished by the centrifugal force caused by the revolution of an orbiting satellite. As a consequence, objects

CONNECTIONS

See also

1 The earth's mass W can be measured by means of a tall balance with two pans on each side. Initially, m1 balances m2 (A) but when m1 is moved to the lower pan it weighs more, being closer to the earth (B). Mass c is used to restore the balance (C). Next, the large mass M is used to make m1 weigh more by gravitational attraction (D). Balance is restored (E) with weight n. If R is the earth's radius d is the distance between m1 and M, then Newton's law states: $\frac{m1 \times M}{d^2} = \frac{n \times W}{R^2}$. The distance between the pans on each side must be great enough to prevent M from having significant gravitational pull on the other masses.



2 In 1798, Henry Cavendish determined the earth's mass using a quartz fibre torsion balance. Two small lead spheres were attracted by two larger lead spheres and from the size of the deflection, the force of attraction between the spheres was calculated. To work out the mass of the earth, Cavendish compared this force with the gravitational pull of the earth on the spheres (that is, their weight).

3 On a roller-coaster, power is needed to pull the car to the top of the highest incline against the gravitational force exerted by the earth. The resulting potential energy of the car is then transformed into kinetic energy of motion as it is allowed to free-wheel down the track. According to the principle of energy conservation, the total kinetic energy at the bottom of an incline should equal the potential energy at the top, the car would be able to travel to the top of another equally high incline with no further application of external power. Some of the potential energy, however, is lost as a result of its conversion to heat energy by friction between the speeding car and the track. Therefore the remaining kinetic energy will allow the car to climb only a smaller vertical height each time it travels downwards. To raise the height of the car again, it would be necessary to apply additional power against the force of gravity.

within the satellite are able to float freely within its confines.

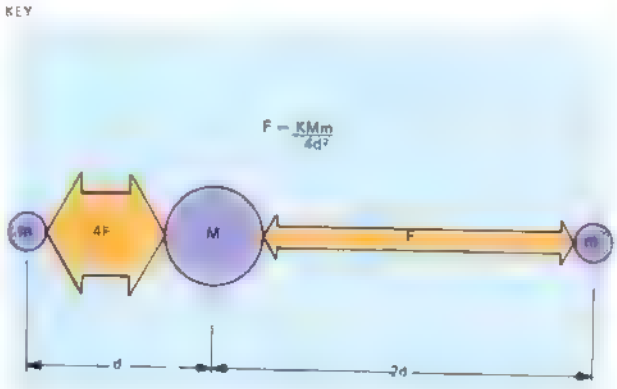
Magnetic force [4] is familiar in connection with the working of a normal compass. This force also acts over large distances, although in this case both attractive and repulsive forces may occur, a fact that is easily tested with two simple bar magnets. The ends, or "poles", of the magnets are distinguished by being called north and south and it is always found that two north or two south poles will not remain in contact, whereas a north and south combination will. This fundamental effect is summarized by the rule: "like magnetic poles repel each other, unlike poles attract".

The different poles are given the descriptive labels north and south for historical reasons – the north pole of a magnet is the one that is always attracted to the North Pole of the earth, though in fact a magnetic south pole must be sited there to achieve this magnetic effect [5]. The inverse-square law determines the magnitude of the magnetic force, although this is now proportional to the product of two magnetic pole strengths that can be positive or negative (instead of two masses that can only be positive, as in the case of the gravitational force).

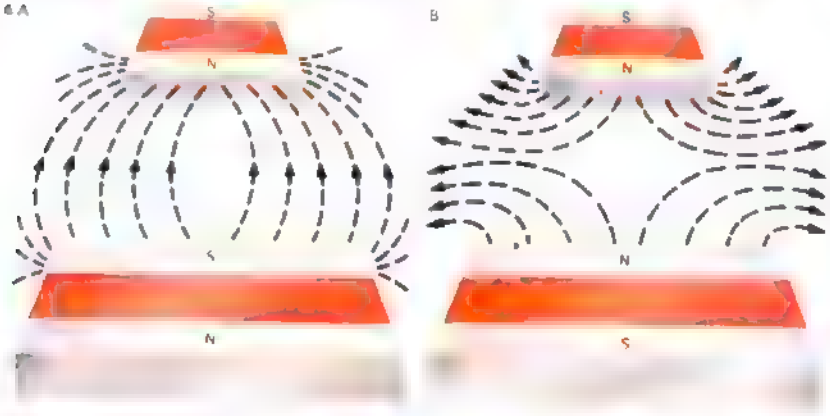
Electric force [6] can also be either attractive or repulsive since its source, an electric charge, can be either positive or negative. According to the type of charge either a positive (attractive) or a negative (repulsive) force is obtained by the inverse-square law.

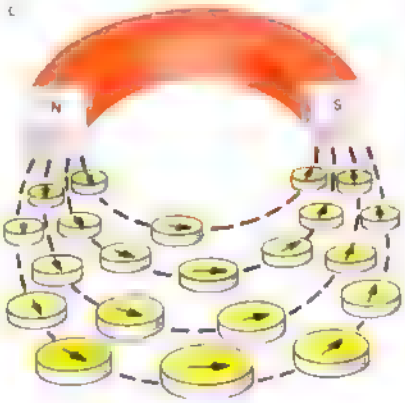
What is an electric charge?

Every electric charge is in fact a multiple of a unit charge that is equal in magnitude to the irreducible charge associated with a single electron, a fact first noted and evaluated in 1909 by the American physicist Robert Millikan (1868–1953). However the dual nature of the electric charge had been known from Greek times and is best demonstrated, as then, by electrostatic effects. If a glass rod is rubbed with silk the two objects attract each other, while two glass rods rubbed in the same way repel each other. Each rod acquires a net positive electric charge, whereas the silk collects excess electrons by the action of friction and becomes negatively charged.



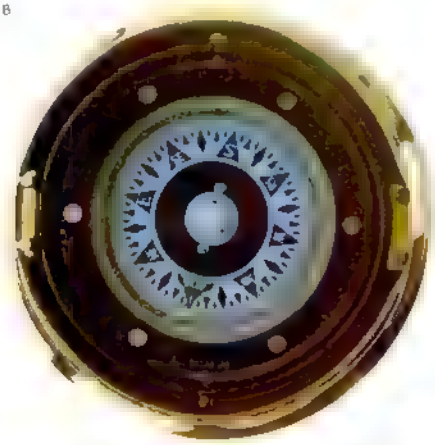
The inverse-square law governs gravitational, electric and magnetic forces. M and m represent the masses of two bodies for the gravitational force, their charge values for the electric force and their pole strengths for the magnetic force. In all cases, d represents the distance between the bodies. K is a constant quantity that has a different value for each of the three forces. Gravitational force is much the weakest, the gravitational attraction for two electrons being about 10^{42} times less than the electric force of repulsion. (Consequently one is aware of gravity only when great masses such as the earth are involved.) The inverse-square law equation thus varies for the three forces according to the value of K.

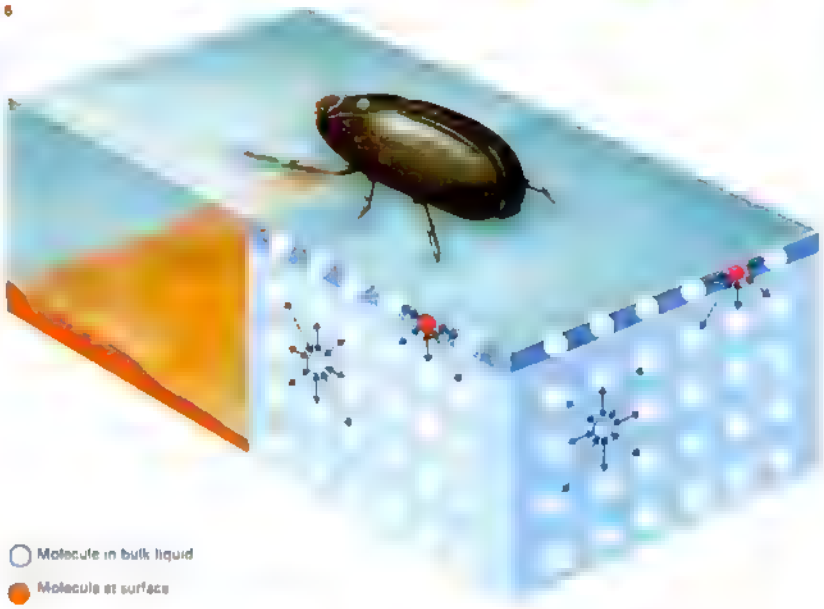




4 A field of force exists all round a magnetic material and the associated lines of force indicate paths along which a unit magnetic north pole would move. Iron filings sprinkled on paper laid over a pair of magnets will display negative [A] or positive [B] lines of force. In a similar way a small compass needle will line up along lines of force with its north pole pointing out the field direction [C].

5 The earth's magnetic field is created by phenomena below the crust as if it had been formed by an immense bar magnet with its south pole approximately aligned with the north geographic pole sited at one end of the earth's rotational axis [A]. The needle of a ship's magnetic compass [B] swings to a position where its ends point to north and south along a line of force of the earth's magnetic field.





6 Electric forces that exist on a molecular and atomic scale within a substance are responsible for the cohesive forces that hold it together and give it shape and strength. Both attractive and repulsive forces exist in a state of equilibrium, the attractive holding the substance together and the repulsive effectively preventing the atoms collapsing in on each other. A striking result of these cohesive forces between atoms is seen in liquids. Within the body of the liquid, any one atom has equal and opposite forces on either side of it in any direction, there is therefore no resultant force. But at the surface there is a net force pulling atoms into the body of the liquid and as a result the surface appears to behave like an invisible elastic "skin" pulled across the liquid. This surface tension effect is used by insects such as water boatmen and mosquito larvae to keep themselves floating on a pond's surface.

Speed and acceleration

Cars, rockets, falling weights and footballs all move under the action of forces. The branch of physics that studies movement, and the forces that produce and influence it, is called dynamics. It was given a firm scientific basis as a result of the work of Isaac Newton (1642–1727), who formulated the fundamental three laws of motion.

Principles of motion

The first law [1] summarizes the principle of inertia – the basic tendency for anything moving to continue moving and for an object at rest to remain at rest. It states that "an object will remain at rest or in motion at constant velocity unless acted on by a force". Once a car is in motion, both it and its passengers carry on moving unless acted on by a force – such as a braking force. A head-on collision may stop the car, but the inertia of the passengers will cause them to fly forwards from their seats. They may be thrown against the windscreen unless held in place by safety belts which exert a restraining force.

As a result of this law, it is apparent that the greater the applied force on an object the greater its change of velocity. Velocity is merely speed in a certain direction and change of velocity in a given time is called acceleration. So the greater the applied force the greater the acceleration. The second law of motion states that, in addition, acceleration is inversely proportional to the mass of the object being moved.

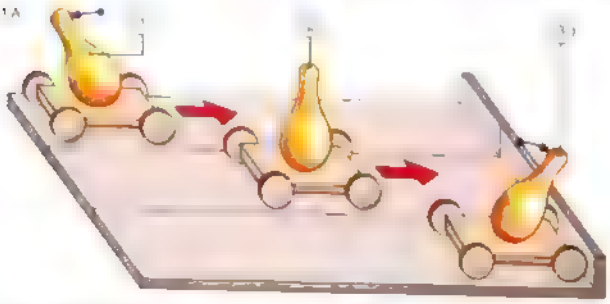
The third law considers the way forces act against each other. If an object rests on a table, the table exerts an upward force equal and opposite to the downward force of the object's weight. The third law generalizes this by saying that "for every applied force there is an equal and opposite reaction". Two spring balances hooked together and pulled in opposite directions register the same force. Another much more spectacular application of this law is the rocket. The force of the expanding gases in a rocket's combustion chamber acts equally in all directions. The forward thrust is produced by the reaction to the displacement of mass that occurs when the burning gases escape through the nozzle of the rocket.

Newton's laws of motion also connect with several other concepts. Thus the second law describes how acceleration (a) is dependent on mass (m) and applied force (F) in the equation $F = ma$. This can be used to calculate the weight of an object, because weight is the force with which a body is attracted towards the centre of the earth. This force equals the product of the mass and the acceleration with which a falling object drops to the ground (called the acceleration due to gravity). Consequently mass and weight are completely different quantities, having different units to describe their magnitudes. Mass is a property of an object due to the quantity of matter it contains; weight, a force which acts on it due to gravity. Unfortunately in the past the same units (pounds, kilogrammes and so on) have been used for both mass and weight; the modern scientific units are the kilogramme (or pound) for mass and the newton (or poundal) for weight.

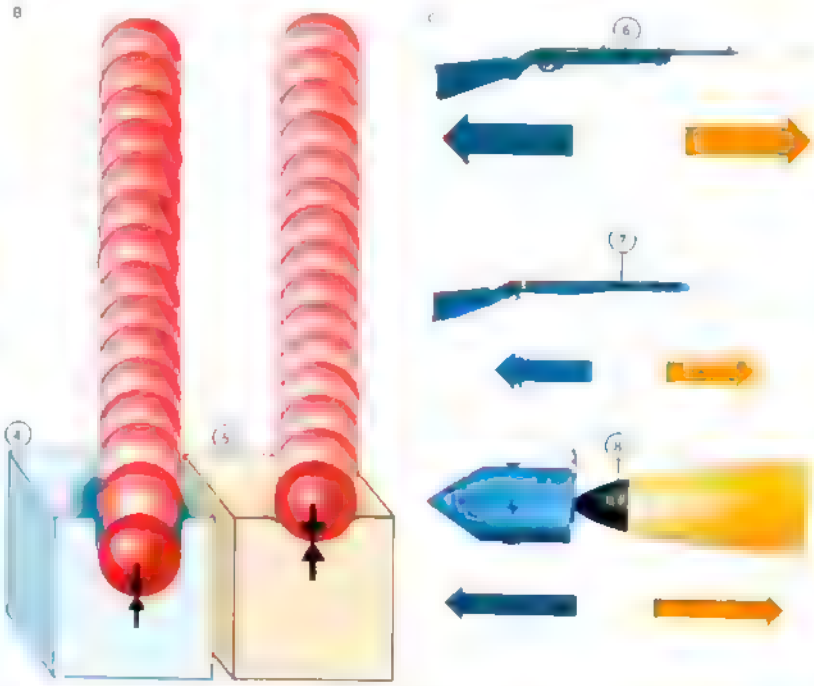
Definition of momentum

The equation of the second law also shows that, because acceleration is the rate of change of velocity, force can be expressed as

CONNECTIONS

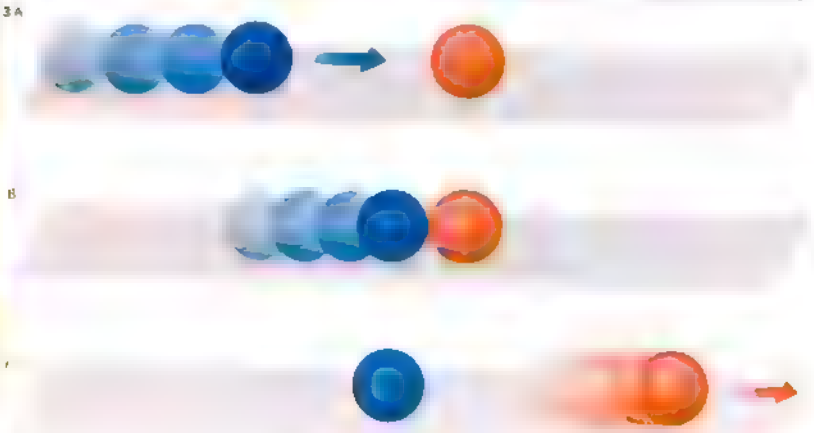


1 Newton's first law describes inertial effects. [A] An object resists being moved from rest by toppling backwards [1], although moving steadily it is undisturbed, as if at rest [2]. When stopped it resists slowing and tends to continue moving [3]. The second law explains that acceleration or deceleration is proportional to the force producing it. [B] A ball falling on to a soft material [4] sinks deeper than into a harder one [5] because the deceleration force is smaller. The third law states there is an equal and opposite reaction to every force. [C] A rifle recoils when fired [6], although the bullet's velocity is much greater. Successive firings cause successive recoils [7]. A rocket ejects gas and moves forwards [8], because of reaction.



2 A heavy gun recoils when firing and an equal and opposite reaction propels the shell forwards. The principle of conservation of momentum states that the total momentum before and after is zero but the shell is given its forward velocity because it is lighter than the gun. The chemical energy stored in the propellant explosive charge is transformed into the kinetic energies of gun and shell.

3 A stationary ball is hit by another of similar mass moving towards it. Before the collision [A] the moving ball has a certain momentum (equal to the product of its mass and its velocity). At the moment of impact [B], this momentum is transferred to the stationary ball, which had no momentum. As a result, the stationary ball moves while the ball that was moving formerly becomes stationary [C]. In all such collisions momentum is conserved – after impact the momentum of the second ball equals that of the first.



the rate of change of the product of mass and velocity. This product is called momentum and can be thought of as a quantity of motion that, for a definite velocity, increases with the object's mass. In effect the momentum indicates the effort needed to move an object or to stop or change the direction of its motion. For instance, a brick carefully placed on someone's foot does not hurt. But the pain caused when the brick is dropped from a height of a metre or so testifies to the effect of momentum.

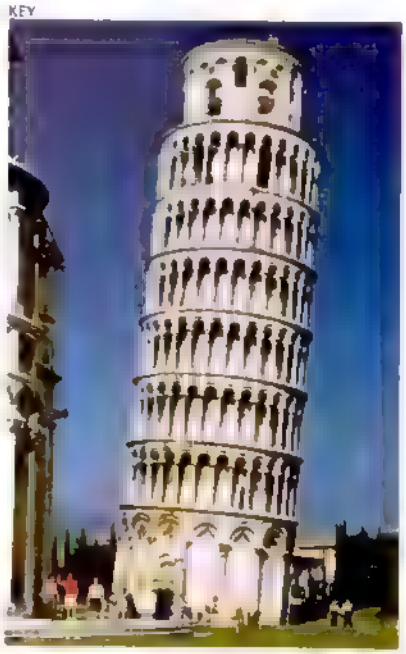
Probably the most important reason for calculating momentum is that it is conserved throughout events involving changes of motion – for example, sudden collisions or explosions. This means that the total momentum before and after such an event stays absolutely constant [3]. Following the event, momentum may be lost through the action of friction reducing the velocities, but during its occurrence this law of conservation of momentum holds exactly.

Another phenomenon occurring during this kind of event is the transfer of energy. It can exist in many forms such as heat, light, sound, chemical and electrical energy. All of these can be transformed into one another – petrol's stored chemical energy is transformed in the combustion engine to mechanical energy for moving a car. If the energy transfers occurring in any process are considered collectively, very careful measurements have shown that energy is never created or destroyed. This is the law of conservation of energy and it implies that no machine can produce a net gain of energy.

Two forms of energy

For the study of dynamics two forms of energy are fundamental – kinetic energy, possessed by objects that are in motion, and potential energy, possessed by objects that are at rest and able to do work by virtue of their position [5]. The pile driver has work done on it to lift it against the earth's gravitational force and then can expend the stored energy when again allowed to fall. In fact, the potential energy is transformed to kinetic energy as the hammer falls and it is the kinetic energy that finally does the work of driving the pile into the ground.

Galileo's legendary experiment, dropping a cannon ball and a pebble from the tower at Pisa, showed that objects of different masses fall to the ground together, proving that the acceleration due to gravity is the same for all objects. With objects of different cross-sectional area, air resistance to downward motion may also be different and prevent the objects hitting the ground together. If resistance equals the gravitational attraction a limiting velocity is reached. Also there will not be a total transformation of potential to kinetic energy as the objects fall because some heat energy is lost by the action of friction with the air.



4 A motorway pile-up can occur after one or two cars stop. Other cars that are unable to stop in time collide with the stationary vehicles and transfer their forward momentum to them, causing a "knock-on" effect which ripples along an ever-extending chain. Conservation of energy controls the transformation of a car's mechanical kinetic energy to wasted heat and the energy imparted to the stationary cars.

5 The change from potential energy (due to an object's position) to kinetic energy (due to its movement) and back again occurs with a bouncing ball (A), a ball rolling in a hemispherical cup (B) and a swinging pendulum (C). Before a ball is dropped it possesses only potential energy. This is converted to kinetic energy as the ball falls and reaches a maximum at the moment of impact – when potential energy is zero. Potential energy is regained as the ball rises to become stationary again at the top of the bounce. The same sequence of transitions occur in the bob at the end of the pendulum and with the ball rolling in its cup.

6 A steam catapult, driven by the pressure of gas built up in the rams, is effectively storing a great quantity of potential energy that, when released, can be transformed to the kinetic energy of movement. As a result, even the large mass of an aircraft can be accelerated to a speed that will allow it to take off from rest in the very short distance available on an aircraft carrier's deck.

7 In 1798 Count Rumford (1753–1814) noticed that, in the boring of cannon barrels, the barrels, borer and metal chips all became hot despite there being no apparent heat source, except for friction. He realized that the mechanical energy needed to turn the barrel against this friction was being converted to heat energy. Subsequently other experimenters, notably James Joule (1818–89), demonstrated the transformations between other forms of energy.

Circular and vibrating motion

If a driver presses the accelerator pedal of his car, the car speeds up – in scientific terms its velocity changes. But even when he goes round a bend at constant speed, the car's velocity also changes. This is because velocity is speed in a certain direction, if either its magnitude or direction alter, velocity changes.

Motion in a circle

The rate of change in velocity is called acceleration. Thus an object which changes direction while travelling at constant speed experiences an acceleration. When a stone is tied to the end of a string and whirled round in a circle at constant speed, the velocity's magnitude is unchanged but the direction alters continuously. If the string were cut at any instant, the stone would fly off along a tangent to the circle (as seen in the sparks flying off a rotating catherine-wheel).

A force must be acting on the rotating stone to produce its acceleration. Here the force is due to the tension in the string and one can feel this force on the hand as the stone is whirled round. It is called the "centripetal force" because it acts towards the centre of the circle. The acceleration it produces is therefore similarly directed towards the centre. For a car moving round a circular track, frictional forces at the ground acting on the wheels provide the centripetal force.

Newton's laws are applicable only to non-accelerating frameworks. For these laws to hold true in the accelerating framework of a rotating system, we must introduce centrifugal force as a fictitious force acting outwards from the orbit centre. The effect of this force is used in the domestic spin-drier to squeeze the water out of wet laundry and, on a larger scale, in the centrifuge machines that subject astronauts and pilots to the high acceleration forces encountered in flight.

Periodic motion

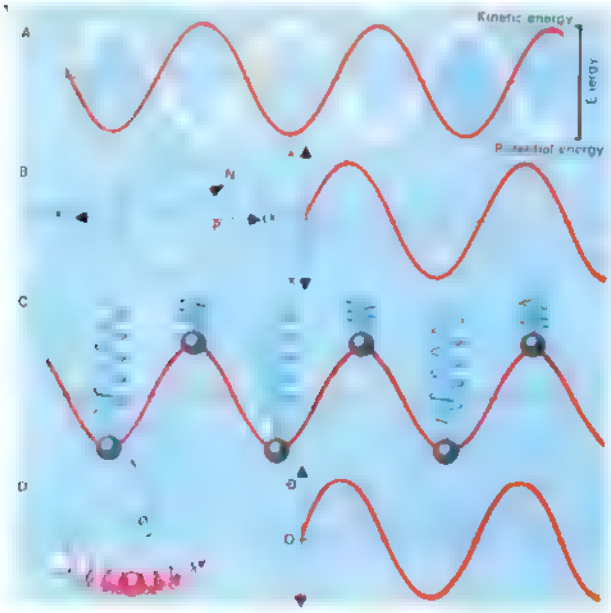
Both centripetal and centrifugal forces depend on the mass (m) of the object and its velocity (v) in the circular motion (circle of radius r). A heavy object needs a greater centripetal force to hold it in orbit and a greater force is also required for high speeds of rotation. Experiment shows that, in addition, the required force (F) is inversely proportional to the radius of motion, and $F = mv^2/r$ where v^2/r gives the magnitude of the centripetal acceleration.

Uniform circular motion is periodic – that is, the events recur over and over again. The time taken for a complete revolution of the object remains constant. This periodic character is further demonstrated by considering how the object's distance from any fixed diameter of the circle varies with time. If a graph of these distances is plotted, the resulting curve is one of a uniformly oscillating displacement [1], resembling a sine wave.

For a swinging pendulum, the force which pulls the bob back through the central vertical position is its weight. It undergoes oscillatory motion with an acceleration that is proportional to its distance from the point of suspension and it is also directed towards that point. Movement of this type is called "simple harmonic motion" and for the pendulum the time taken for a complete cycle of forwards and backwards swing (the oscillation "period" t) is proportional to the square root of its length (l). The number of cycles completed per second is called the "fre

CONNECTIONS

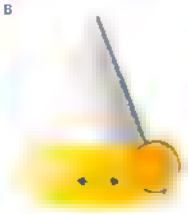
See also

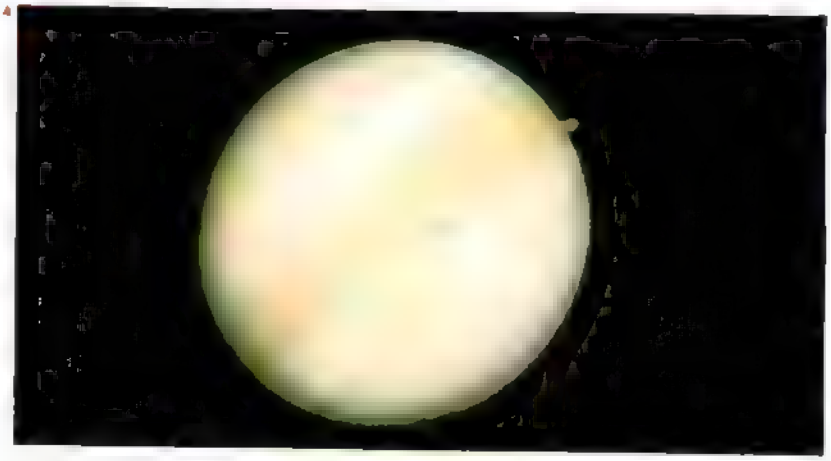


1 All periodic motion involves the continuous interchange of kinetic and potential energy, as shown in the graph (A). Simple harmonic motion (SHM) is one form of periodic motion that is characterized by the shape of a sine wave (B). The point (P) on the diameter of the circle around which N is moving is an example of SHM. The mass on a spring (C) performs linear SHM and a pendulum (D) performs angular SHM.

2 Circular motion of the "aeroplanes" produces an apparent centrifugal force that lifts the planes off the ground. The equal and opposite centripetal force is provided by tension in the arms.

3 The constant periodic time of oscillation of a pendulum is used to control clocks, especially case clocks (A). The periodic time needed to drive the escapement correctly can be exactly matched by choosing a pendulum (B) of correct length. It can be calculated from the equation $t = 2\pi\sqrt{l/g}$ where t is the oscillation period, l is the length and g is the acceleration due to the earth's attraction – gravity.





4 The moons of Jupiter, as well as all other natural and artificial satellites, move in orbits around their mother planet at great speeds. The force of gravity between them provides the centripetal force towards the centre of motion that keeps a moon in orbit and produces its centripetal acceleration. As the speed of rotation remains almost constant, it is the continuously changing direction of the moon which implies acceleration. If the inward gravitational force were removed the moon would continue in a straight line hurtling out into space. The orbits are not exactly circular, so that the speed of rotation does not remain completely constant. The associated forces are analogous to those of a system performing circular motion.

quency" of oscillation (measured in hertz)

The displacement in all types of simple harmonic motion rises and falls like a sine wave. Many other wave motions behave in this way. If a long rope is clamped at one end and its free end is whipped, a wave-like disturbance travels along the rope and the amplitude of the rope's displacement at any point from the fixed end is described by the waveform shown in illustration 6.

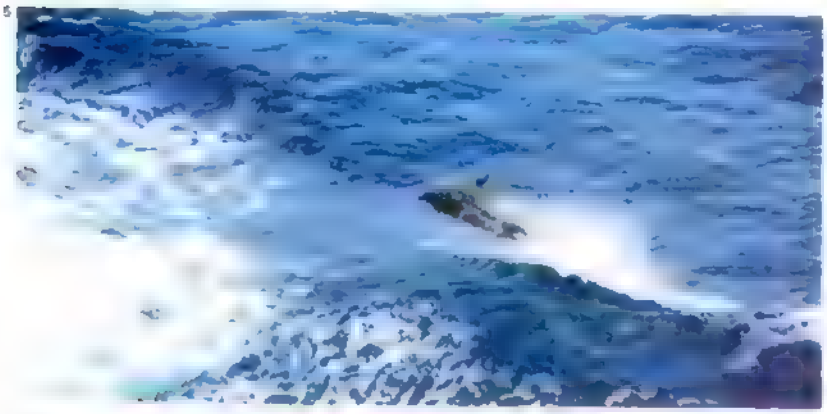
Moving and standing waves

The characteristics of this wave picture describe amplitude changes for the plane waves that move through the sea, the circular waves (ripples) that spread from the point at which a stone is dropped in a pond, the air pressure waves of sound and the electromagnetic waves of radio and light (which are distinguished by their wavelengths). For all these waveforms, energy is transmitted in the direction of the wave motion. Actual vibrations of the medium in which they travel may occur in the same direction, producing "longitudinal waves" such as those of sound, or in a direction at right-angles to the motion, producing the transverse waves of all the other examples. In water, a floating object merely bobs up and down as the waves pass, and it does not move in the direction of the wave motion.

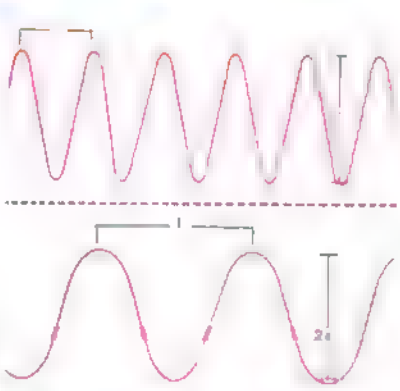
If a string is clamped at both ends and then plucked, it still vibrates but in this case the wave appears to stay in the same place – a stationary, or "standing", wave has been produced as the ends are fixed and allow no forward motion. The string vibrates at its natural frequency, which is inversely proportional to its length and directly proportional to $\sqrt{T/m}$ where T is its tension and m is its mass per unit length.

This state of vibration is called "resonance" because it occurs with the natural frequency of the string. Blowing into the end of an organ pipe makes the air in it resonate at a natural frequency that depends on the pipe's length. Similarly, when struck, a tuning fork vibrates with one particular frequency. There are many useful applications of resonance, as in tuning circuits on radio receivers [8], and in musical instruments – perhaps the most satisfying application of this effect.

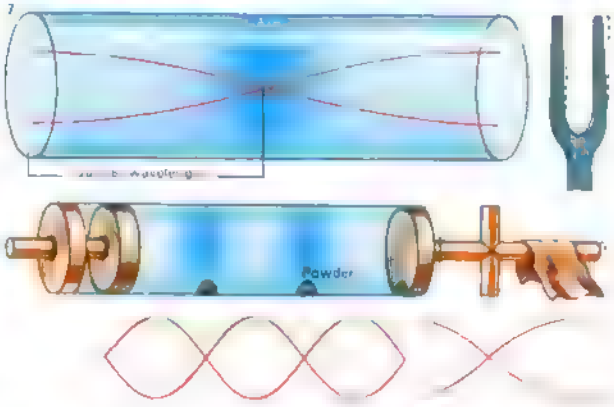
Many fairground rides depend on circular motion and the resulting centrifugal forces for the excitement they can provide. They all use different variations of this type of motion, ranging from the slow revolutions of the merry-go-round to the much faster and more complex movements of the aeroplane "spinners". The magnitude of the centrifugal force depends on the mass of the circling object. On this "chairoplane" the heavier children experience greater centrifugal force than the lighter ones, but do not fly higher because the force is balanced by their greater weight. In ultra high-speed centrifuges small molecules can be separated according to their different mass.



5 Sea waves move with a transverse wave motion that transmits energy in their direction of motion, while floating objects move only vertically up and down as the waves pass them.

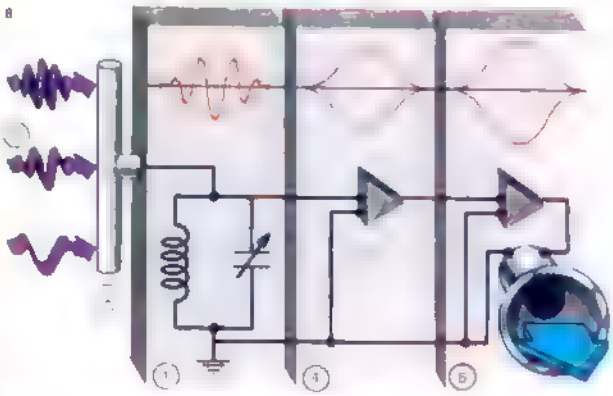


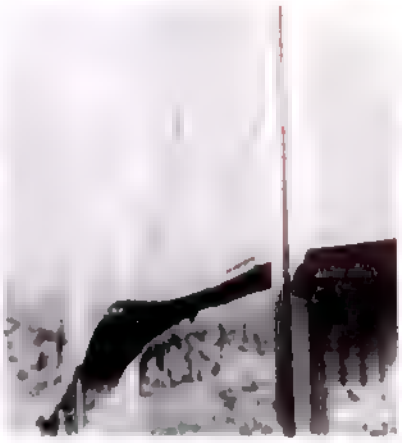
6 The forward velocity (v) of a wave of frequency (f) and wavelength (l) is given by $v = fl$. Such waves can be drawn as sine waves. Two waves can have the same amplitude (a) even though the wavelengths differ.



7 The air inside a pipe will vibrate at a natural frequency that depends on the pipe's length. If the external impulse has this frequency resonance occurs. Powder in the lower tube shows the wave.

8 A circuit [1], with the AC source from an aerial [2], is "tuned" so that its natural frequency matches one of the incoming radio signals [3]. Resonance excites it to oscillate and work the radio [4, 5].





9 Freak winds can begin to produce oscillations in very large and heavy structures. If they continue, the oscillations may gradually increase in amplitude and if their frequency equals the natural vibration frequency of the structure resonance may occur, leading to catastrophic break-up of the structure. This has happened with bridges (the most renowned probably being the Tacoma Narrows Suspension Bridge, USA, shown here) and tall structures such as skyscrapers.

Pressure and flow

The branch of physics that deals with the forces and pressures that act within liquids and gases is called hydrostatics. It also examines how these forces affect any surface they touch and what uses they can be put to. It includes, for example, problems ranging from deep-sea diving to aeroplane altimeters and from floating and sinking to the design of hydraulic lifts and other machinery.

Archimedes' principle

Hydrostatics explains what happens to an object immersed in a fluid. Both liquids and gases are fluids and each is able to exert or transmit a force. If a cork is pushed below a liquid surface and then released, it immediately bobs to the surface. The cork has experienced an upward force, called the "upthrust", due to the liquid, and it is this that keeps the cork floating on the surface. In exactly the same way, an upthrust must act on a floating balloon, although here it is produced by a gas rather than a liquid.

It was the Greek scientist Archimedes (287–212 BC) who first quantified this fact by stating that "when an object is totally or partially immersed in a fluid the upthrust on it is equal to the weight of fluid displaced". Using Archimedes' principle [1] the magnitude of the upthrust can always be found and it makes itself apparent as a loss of weight of the object.

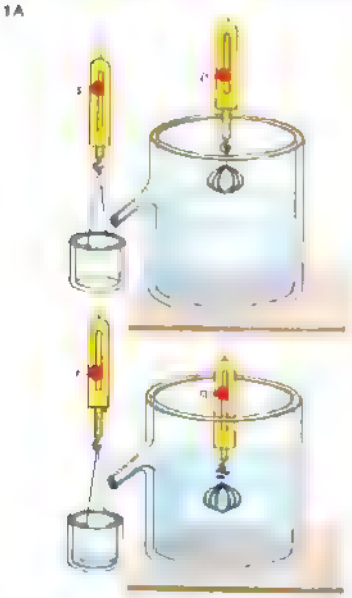
For this reason, a floating object has its weight exactly balanced by the upthrust. But if the object is too dense, the upthrust may not be sufficient to counterbalance its weight and the object sinks. This principle of flotation is used directly by a hydrometer, an instrument for measuring liquid densities. A hydrometer floats at a level that depends on the weight of a liquid it displaces and, since the submerged volume is known, the density of the liquid can be calculated.

Internal forces in fluids

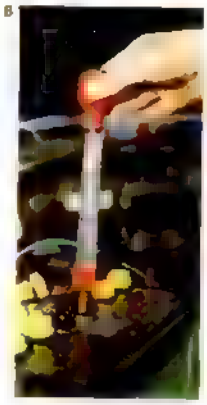
As well as being able to exert a force such as an upthrust, a fluid can produce an internal force at any depth due to the weight of the fluid above it. It is normal to measure this effect as the force per unit area, or pressure, developed by the fluid's weight; this increases with depth (in both liquids and gases). Thus the greater water pressure at the bottom of a dam requires a form of construction in which strength also increases with depth – a triangular cross-section getting thicker with depth is generally used. For a similar reason, the suit of a deep-sea diver must contain a jacket of compressed air whose pressure counteracts the external water pressure, so that he can breathe without his muscles having to expand his chest against this pressure.

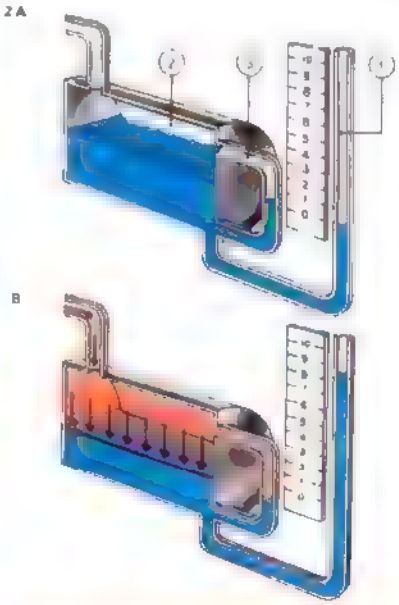
Another example of a fluid is the earth's atmosphere. The weight of air produces a pressure at the surface, commonly called atmospheric pressure, which is about 14.7 pounds per square inch at sea-level. In other unit systems it is equal to 760mm of mercury, 101,325 newtons/m^2 or 1.013 millibars. Variations in this pressure affect the weather and are caused by atmospheric disturbances. They are measured by the common barometer. The simplest form of barometer measures the height of a column of liquid, often mercury, supported by the atmospheric pressure. An aneroid barometer transforms the effect of pressure on a thin-walled metal

CONNECTIONS

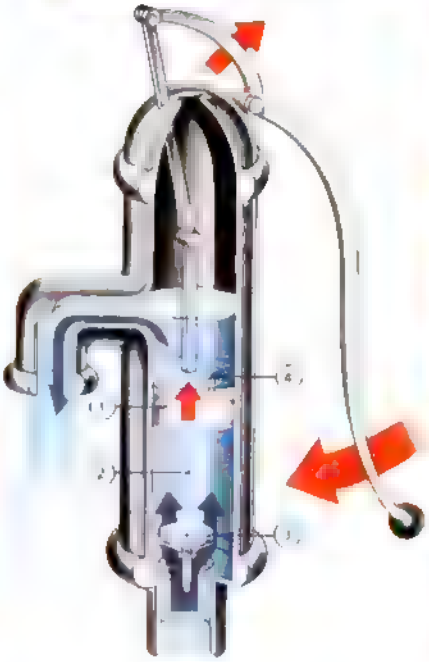


1 Archimedes' principle states that for an object immersed in a fluid the upthrust on it equals the weight of fluid that its volume displaces [A]. The round object suspended on the balance has weight p in air and lesser weight q when immersed in a liquid. The difference in its weight is equal to the upthrust on it in the liquid, that is, to the weight of displaced water in the beaker, which equals r minus s and which in turn equals p minus q. An application of Archimedes' principle is the hydrometer [B], which is used to test the condition of a car battery by checking the acid density. The tube is immersed in the battery acid and the bulb is squeezed to expel air; it is then released to suck acid into the main stem. The hydrometer then floats in the acid, at a depth that depends on the liquid's density.





2 The manometer is a U-shaped liquid column gauge used to measure differences in fluid pressure. The "well-type" has one column [1] of relatively small diameter whereas the second that is wider acts as a reservoir [2]. The difference in columns ensures that the level in the reservoir does not change much with pressure, but that in the small diameter column does, so making for accuracy in reading these variations. Small positional adjustments of the scale with the help of the level indicator [3] compensate for the small reservoir changes. At first [A] both columns are at equal levels. With the reservoir pressurized [B] the new level shows pressure.



3 In the lift-pump, the piston [1] is moved upwards by the downwards stroke of the lever, producing a partial vacuum in the cylinder [2]. The piston valve [4] is kept closed by the water already in the pump chamber. Water is then forced into the cylinder by atmospheric pressure acting on the surface of the water outside the pump. This water passes through the open valve [3] filling the upper chamber. On the upward lever stroke the piston moves down, valve 3 closes and valve 4 opens to allow the piston to move through the trapped water. On the next downward stroke this water is lifted out of the pump and the cycle begins again. In theory, atmospheric pressure "lifts" up to 10m (34ft).

4 A water cannon has water pumped into it under great pressure and forced through an exit nozzle of relatively small diameter. As a result a great force propels the water forwards (pressure is the force per unit area) and accelerates it to a high velocity. This means that the water acquires considerable momentum and kinetic energy which can be used to cut relatively soft china clay from a quarry wall. The great power of the water cannon is also often used to clean the outside walls of buildings. In this case, an abrasive in the form of a powder can be put into the water to increase the corrosive effect of the water jet as it strikes the surface to be cleaned, removing accumulated dirt.

cylinder into the mechanical movement of a needle moving across a calibrated dial. This form is used as an altimeter in many aircraft.

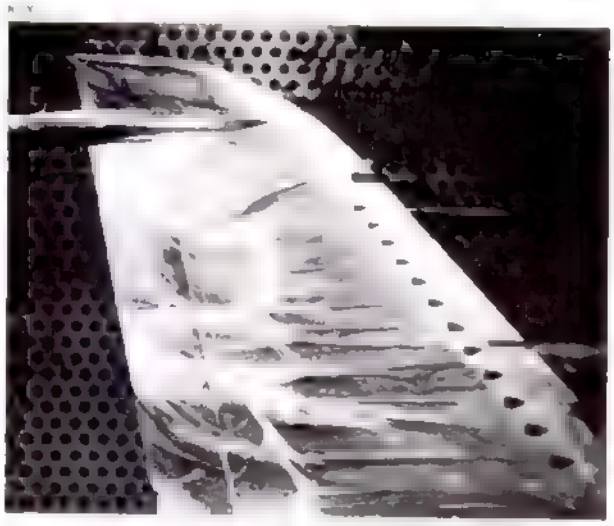
The possibility of supporting a column of liquid by gas pressure is also used in a manometer [2]. This generally consists of a U-shaped glass tube containing a liquid that moves round the U-bend by an amount depending on the difference between two pressures applied at each end. The same principle is employed in the common pump that lifts water from a well. It "taps" a column of water supported by atmospheric pressure acting on the surface of the water source. The height of this column can theoretically be 10.36m (34ft), although in practice the so-called lift-pump [3] can raise water from a depth of only about 8.5m (28ft).

External pressure can be used to move a fluid, but its own internal pressure can also be most effectively employed. A liquid is virtually incompressible, so that pressure developed at one point is transmitted equally in all directions. This fact can be utilized in a hydraulic press or jack to exert very large forces. A force acting on a very small area produces an enormous pressure (force per unit area) which can be transformed into a much greater force acting over the large area of the hydraulic jack ram [5], since the pressure is constant throughout a liquid.

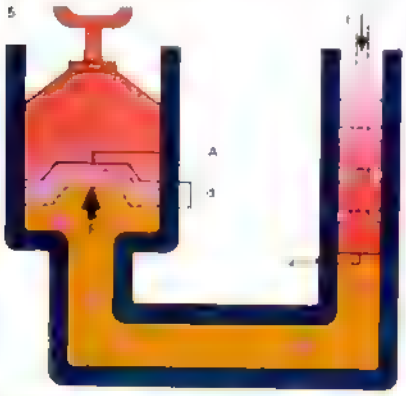
The study of hydrodynamics

All these effects use the static properties of fluids. But by definition a fluid is something that flows and the properties resulting from motion are described by hydrodynamics.

Motion changes the pressure within a fluid, and this can be difficult to predict accurately. The flow can be either smooth (streamlined) or turbulent, when the fluid is broken up into eddies, it is then harder to calculate the pressure at different points within the fluid. The Swiss scientist Daniel Bernoulli (1700–82) first noticed that pressure decreases as fluid velocity increases (although this holds accurately only for streamlined flow) and this principle creates the necessary lift on an aircraft's wing [Key]. The wing shape is arranged to produce a greater flow velocity above than below and so there is a net upward pressure or lift.

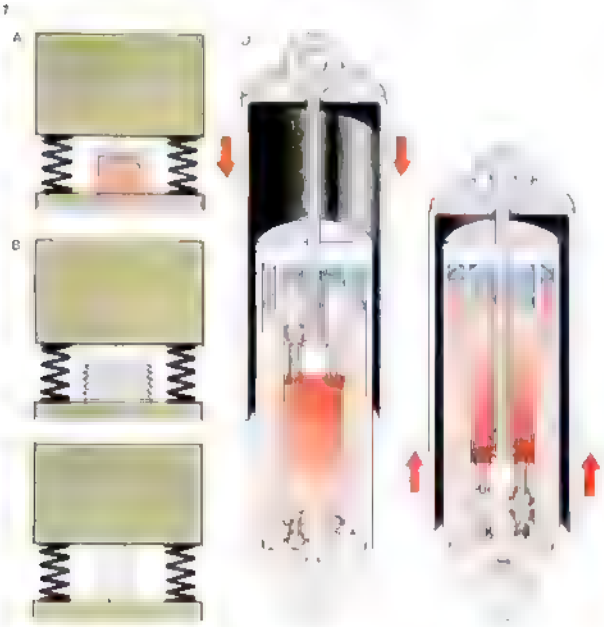


The way air flows around objects can be studied using a wind tunnel. Thus the aerodynamic properties of aircraft wing shapes, or of complete planes and cars, for example, can be investigated. Tiny plastic spheres are injected into the airstream so that the flow patterns can be photographed and streamlined designs created.

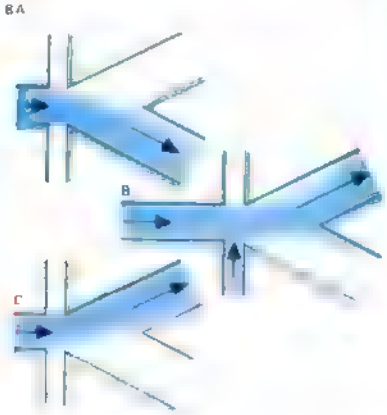


5 An hydraulic lift works on the principle that an incompressible liquid transmits pressure equally in all directions. The small force f acting on area a produces pressure f/a which is converted at the much larger area A to the correspondingly larger force F so that F/A = f/a. The larger force can be used to move great weights. But the weight moves through only the small distance d.

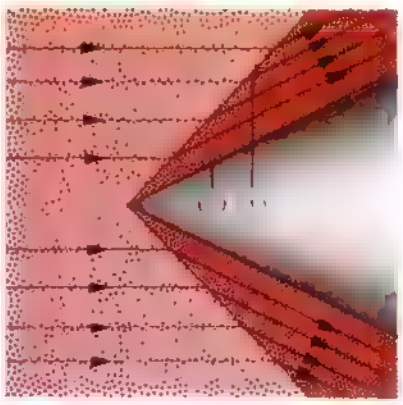
6 The principle of hydraulic lift has many applications, the most familiar of which is probably the car lift used by motor mechanics in garages. Other types of hydraulic machinery have many other applications in agriculture and industry, as in this hydraulic hoist used by an engineer to rig or maintain overhead telephone and electric power lines. If such machines worked exactly as described in illustration 5, the pressurizing force would have to move several metres to produce a small movement of the actuating piston. The hydraulic fluid is therefore pressurized by a pump, which transfers the necessary pressure.



7 A shock absorber is used to reduce or dampen oscillating motions. Several damping systems can be employed – oil damping is shown in A, air damping in B and friction damping in C. The most common type is the oil damper D, which is used especially in road vehicles. It consists of a plunger that on its downward stroke allows oil to pass through a small valve in the piston. The viscosity of the oil and the size of the valve determine the damping characteristics, that is, the manner and speed with which the oscillation is reduced. For rapid loading conditions, where speed of application makes damping solid, a second valve channels oil into a reserve.



8 A stream of gas tends to adhere to any adjacent solid surface, a phenomenon called the Coanda wall attachment effect. A gas stream that is free to enter two identical channels (A) chooses one of them by adhering only to its walls. A small side stream (B) can then deflect it to other output channels (C) where it remains even when the disturbance ceases. This is the fluidics analogue of an electric switch.



9 When the velocity of a gas is low compared to that of sound its flow can be described without considering the compressibility. Above 0.3 of the speed of sound (Mach 1) a gas is compressed when it meets a solid and there is a consequent temperature change. Above Mach 1 compression occurs abruptly and a shock wave forms. The diagram shows a Mach 4 shock wave (1) at an aircraft nose-cone (2). At the shock front the direction of flow changes and the gas becomes more dense.

What is sound?

Sound is energy and, like other forms of energy, can be useful to man. The vast range of expression that characterizes both speech and music makes sound a highly efficient medium of communication and ultrasound – sound above the hearing range of man – is used like radar by bats, dolphins and whales. Even loud sounds do not represent a great deal of energy. A symphony orchestra playing as loudly as possible involves sound energy equivalent to the light and heat energy from only a low-powered electric lamp. Our hearing sense is more easily saturated (in energy terms) than our visual sense.

How sound is produced

Sound is a particular form of kinetic energy (energy of motion) produced when any object vibrates. Vibration is the cause of all sounds, although generally it is not visible. The sound of a car crash booms out as the surfaces of the two colliding vehicles vibrate with the force of the collision; music comes from a radio as the cone of a loudspeaker vibrates; and talking and singing result from vibrations of the vocal cords in the larynx.

As an object vibrates it sets the air molecules around it vibrating. The vibrations move out through the air, forming a sound wave, but the air does not move along with the wave. Where the air molecules gather together a region of higher pressure (compression) forms. Where they move apart a region of lower pressure (rarefaction) occurs. A succession of compressions and rarefactions move through the air as the sound wave passes. At the ear, they set the eardrum vibrating and we hear sound.

If a surface vibrates more strongly the pressure difference between the compressions and rarefactions is greater and the sound is loud [1]. The frequency of vibrations affects the pitch, or note, of the sound. If it is fast, the compressions and rarefactions are close together and the pitch is high. A slower speed of vibration causes the compressions and rarefactions to be farther apart and the sound is lower in pitch.

A sound wave moves out from its source in all directions, travelling at a speed of 331m (1,087ft) per second or 1,194km/h (741mph) in air at sea-level. The speed is slower at high altitudes as air is less dense there and faster in water and metal because these substances are more elastic than air and transmit vibrations more rapidly. Sound cannot move through a vacuum because there are no gas molecules to vibrate and transmit the sound.

Like other waves of energy, sound normally travels in straight lines, but sound can turn corners. It is reflected whenever it strikes a surface such as a wall [2] or floor and is diffracted or spreads out as it passes through an opening such as a window [3].

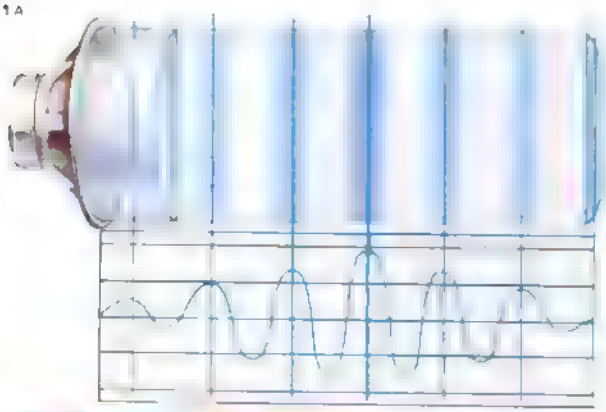
Dynamics, pitch and frequency

The loudness of a sound can be measured with a decibel meter and the result given as a number of decibels (dB). Strictly, the meter measures the intensity of the sound, which is related to the pressure differences in the sound wave. The scale is logarithmic – an increase of 10dB is produced by 10 times the intensity. Loudness varies with the cube root of intensity, so that a sound 10dB greater sounds about twice as loud. The human ear does not hear all frequencies of sound in the

CONNECTIONS

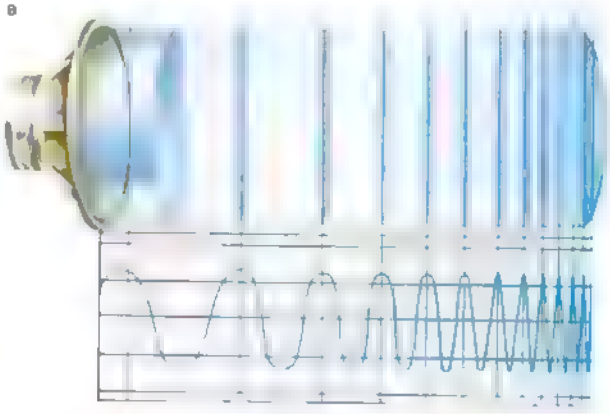
See also

1 A



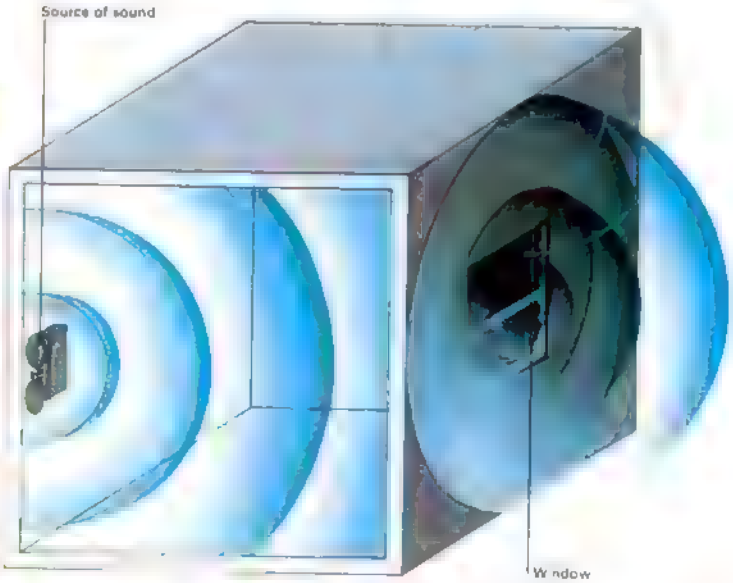
1 **A sound wave** consists of pressure differences, shown as dark and light bands [A]. The curve shows how pressure changes with time. This wave has a constant frequency (a single note) but decreases and increases in intensity. It would have a "wah" sound. A falling note [B] would be heard as this sound wave passes. The frequency decreases as the note becomes lower, but the intensity remains the same.

B



2 **The Whispering Gallery** in the dome of St Paul's Cathedral in London is renowned for its acoustics. A sound whispered against the wall on one side of the gallery can be heard clearly on the other side. Being circular in shape and made of stone, the walls reflect the sound of the whisper all round the gallery and concentrate the sound at the opposite side, 32.6m (107ft) away. Normally, a whisper would be inaudible at such a distance.

3



3 **Both reflection and diffraction** enable a sound to be heard even though the person or object producing the sound is hidden from view. Sound is reflected from surfaces such as walls, floors and ceilings and, more over, undergoes diffraction at an opening such as a door or window. As they pass the edges of the opening the sound waves spread out and hence the opening appears to be the source of the sound.

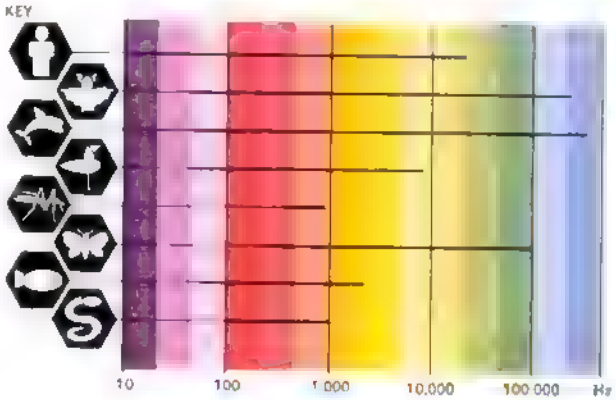
same way and a low sound is perceived as being less loud than a high sound of the same intensity

The number of compressions that pass in every second is called the frequency of the sound wave and is measured in hertz (Hz), equal to cycles per second. This scale is not logarithmic and a note of 440Hz (the A above middle C in music) sounds twice as high as, or an octave above, one of 220Hz (the A below middle C). In other words, the higher the frequency, the higher the pitch

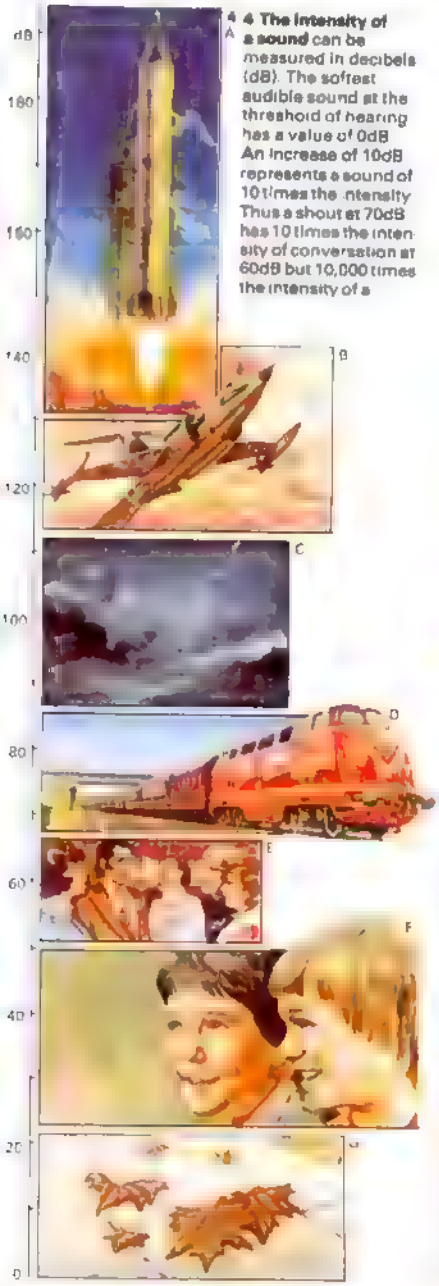
Noise and acoustics

Noise does not have any particular pitch and covers a wide frequency range [4]. Very loud noise is dangerous as well as a nuisance, because continuous exposure to sound of more than 100dB – the levels produced by jet aircraft [5] and machines in many factories – soon results in a permanent reduction in hearing ability. Low-frequency noises are particularly hazardous because they do not seem to be as loud as higher sounds and tests have shown that very high levels of low-frequency sound and infrasound (sound below the hearing range of the ear) quickly result in vertigo, nausea and other physical effects; military scientists have even conducted experiments with infrasound as a potential weapon

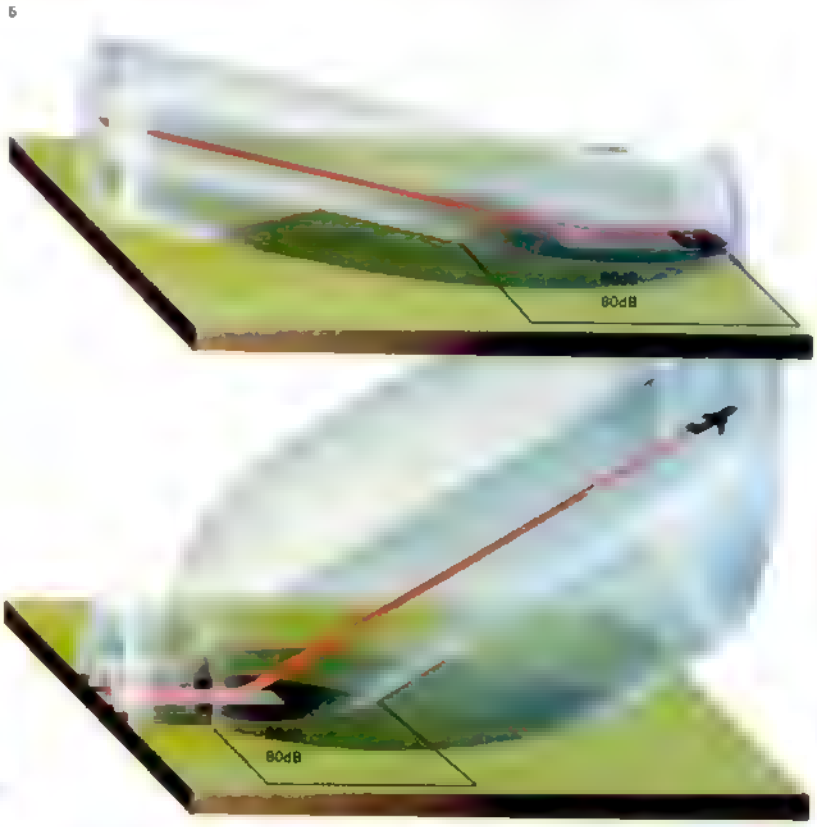
Acoustic engineers work to reduce noise and improve sound in many ways. A consideration of acoustics in the design of a machine such as a jet engine can reduce the amount of noise it makes. Buildings can also be designed to prevent the transmission of sound through them. A steel framework tends to distribute sound throughout a building, but the use of soft sound-absorbing materials in and on floors, walls and ceilings prevents sound from getting into and out of rooms. In concert halls the reflection of sound inside the hall is rigorously controlled to provide an exact amount of echo and give the best quality sound [6]. This may be assisted by electronic amplification, although very loud music loses clarity in a concert hall. Some recording studios have completely absorbent walls to remove all echo and ensure total clarity whatever the type of music being played



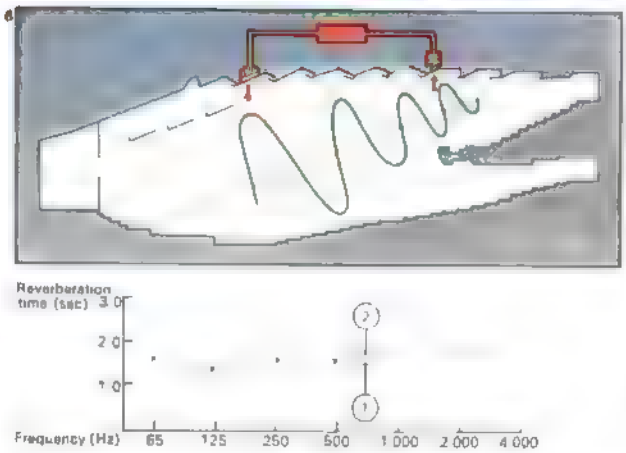
The range of hearing varies widely in man and other animals. Birds and man have fairly similar hearing ranges and both use sound to communicate. Bats and dolphins are sensitive to ultrasound (beyond human hearing), which they use to avoid obstacles and to find their prey by echo location. There is good evidence that dolphins and other whales communicate by means of ultrasound. Night moths make use of ultrasound to avoid predators. Mosquitoes hear a narrow range of sound, corresponding to their own buzzing. The range heard by fish is also extremely small



4 The intensity of a sound can be measured in decibels (dB). The softest audible sound at the threshold of hearing has a value of 0dB. An increase of 10dB represents a sound of 10 times the intensity. Thus a shout at 70dB has 10 times the intensity of conversation at 60dB but 10,000 times the intensity of a whisper at 30dB. The chart shows the intensity of some sounds near their sources. At 140dB sound causes pain. These fairly common sounds illustrate the range that can be heard by human beings: [A] space rocket at lift-off, 140–190dB, [B] a jet aircraft on take-off, 110–140dB, [C] thunder, 90–110dB, [D] a train, 65–90dB, [E] loud conversation, 50–65dB, [F] quiet conversation, 20–50 dB, and [G] a rustling of dry autumn leaves, 0–10dB



5 Noise is a hazard, particularly near an aircraft, and airport personnel who work on the tarmac may wear earmuffs for protection. Farther away, noise is a severe nuisance but not physically dangerous. Near an airport noise levels are about 90dB – sufficient to drown all conversation – and for miles around noise levels may be 60dB – roughly that of heavy traffic. The problem is worse when aircraft are landing than at take-off. A landing aircraft flies nearer the ground for a longer time and its noise consists of disturbing high pitched whines. After take-off an aircraft climbs rapidly and leaves a smaller area affected, also its sound is more of a rumble. The new generation of jet aircraft – supersonic airliners excepted – have quieter engines and may affect an area only a tenth the size of that disturbed by older jet aircraft



6 The acoustics of a London concert hall were altered when it was discovered that the reverberation time – the time taken for the sounds made on the stage to die away in the hall – was too short for the lower frequencies. Electronically amplified resonators were placed in the ceiling to add echo to the hall. The graph shows the reverberation time before [1] and after [2]. The result was a more balanced and pleasing sound throughout the hall.

Musical sounds

Why should one musical instrument sound so different from another? Instruments are played in various ways; some are struck, some are blown, while others are bowed or plucked to produce many kinds of sounds. But what is different about the sound itself?

Frequency and pitch

Every instrument produces a sound by making something vibrate and the frequency of the vibration is related to the pitch of the note produced. If the vibration is more rapid, the number of vibrations in the sound wave that reaches the ear (the frequency of the wave) is greater and the pitch is higher or more to the treble. If the frequency is less, the pitch will be lower or more to the bass.

The frequency of a sound wave (number of vibrations per second) is measured in hertz (Hz). The audible range of frequencies for most human beings lies between 20 and 20,000Hz. But some animals, bats and dogs for example, hear over a far wider range.

Every instrument produces a certain set of notes within a particular range of pitch. But each note is in fact a combination of many more notes. The pitch of the main note heard by the ear is called the fundamental, and above it every instrument also produces a group of higher-pitched notes called harmonics. The harmonics are produced because the vibrating object making the sound vibrates at several frequencies at once and the extra frequencies are simple multiples of the fundamental frequency.

These higher notes can sometimes be produced deliberately on certain instruments – on brass instruments by blowing harder and on string instruments by a particular method of fingering – but normally they are not heard individually. If they were, each note on an instrument would sound like a vast chord. Instead, all the harmonics combine with the fundamental note to produce a complex waveform. Each instrument produces its own particular waveform because the relative intensity of the harmonics is different. The modern music synthesizer works by producing several waveforms of basic shapes – a sine wave, a saw-toothed wave and a square wave – and then combining them to make all kinds of sounds.

Not all instruments produce a note of definite pitch. Several, such as drums and cymbals, produce noise, which consists of a wide range of frequencies without any particular dominant frequency.

The effect of volume

Volume, or the degree of loudness, is another quality of musical sound. Music employs contrasts of volume on a large time scale for dramatic effect, but on a small time scale the change of volume at the beginning of a note is essential to the quality of a sound. The starting characteristics [3], called transients, determine whether a note begins quickly or takes some time to build up, transients are complex and involve changes in the waveform as well as in volume as the instrument begins to sound. Transients are vital to recognition; if the transients are removed from a recording of an oboe, for example, the character of its sound changes until it sounds more like a mouth organ.

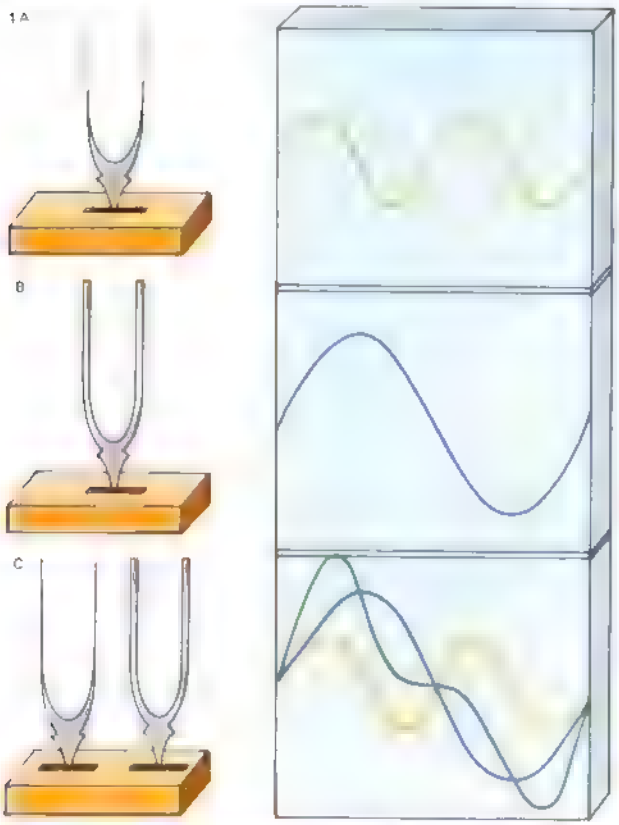
Two other qualities often present in a musical sound are echo and vibrato. Echo is often believed to improve music, giving it a

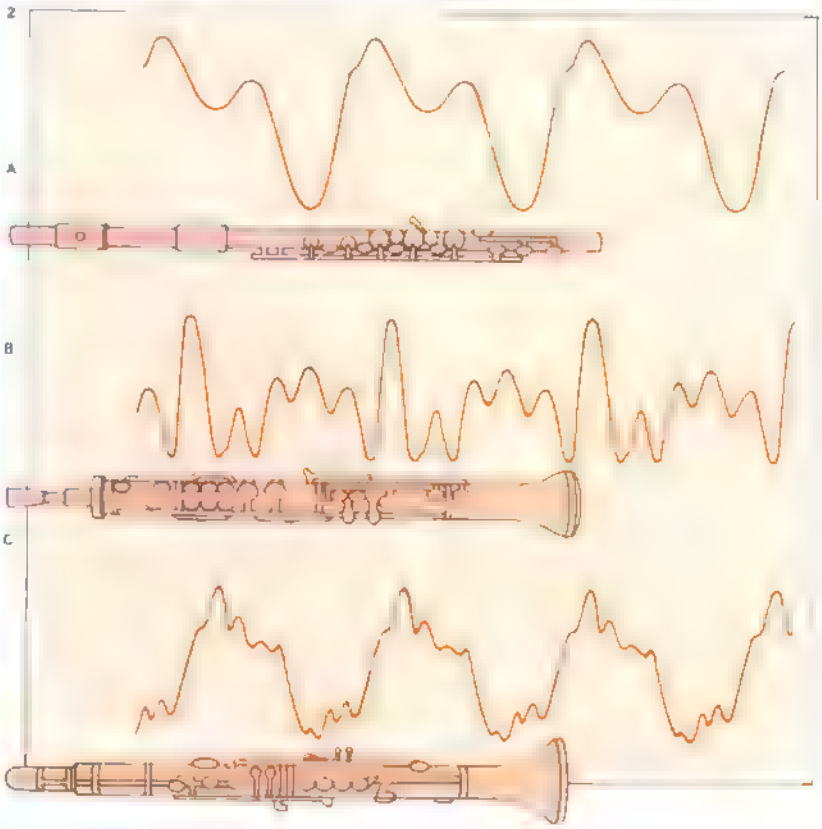
CONNECTIONS

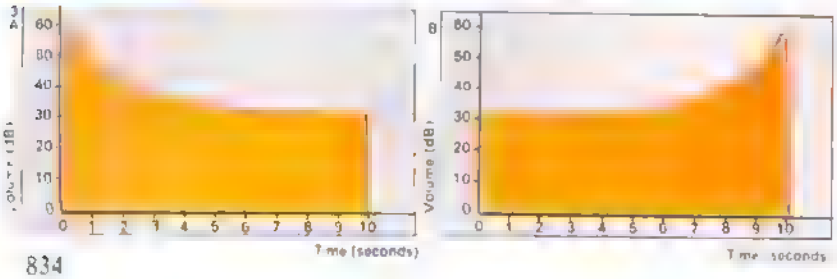
See also
What is sound?

1 Sound waves combine and the waveforms show that the first tuning fork [A] has twice the frequency of the second one [B]. The combination [C] produces a sound equal to the differences in frequencies and the altered shape of the waveform (green curve) shows that it has changed in tone. Two waves that are only slightly different in frequency combine to give a slow beating (pulsing) sound.

2 The waveforms of a flute [A], oboe [B] and clarinet [C] show the differences in tone of the instruments. The flute's rounded waveform indicates a gentle, fluid sound. The clarinet wave has a similar shape with "jinks" reflecting the instrument's more reedy sound. The jagged waveform of the oboe shows that its sound is even more reedy. Every note of an instrument has a fundamental frequency but also vibrates at frequencies that are simple multiples of this, to give a range of harmonics above the fundamental. For example, G below middle C has a fundamental of 196Hz with harmonics of 392Hz, 588Hz etc. Harmonics collectively colour the basic note by combining with it to give a complex waveform. The relative intensity of the harmonics shows that the fundamental is less intense than some of the harmonics. But it is reinforced because the harmonics interact to give notes with frequencies equal to their differences.







3 The start of a sound must be heard if it is to be recognized. The graph shows that a piano note, for example, reaches its peak volume very soon after it has been struck [A]. It then begins to fall away with great rapidity at first, before falling away slowly for several seconds. This is what gives the instrument its "attack". The same note produced on a gong takes a comparatively long time to build up and, although of the same pitch, sounds completely different. The same piano note played backwards on a tape recording [B] slowly increases in volume and then suddenly stops. Shorn of its start, the sound is totally unlike a piano and rather like an organ.

more rounded sound and it is produced by the reflections of sound from the walls of a concert hall or added artificially to recordings. Vibrato is a slight wobble in pitch that many musicians like to use; a violinist moves his left wrist to and fro to produce vibrato.

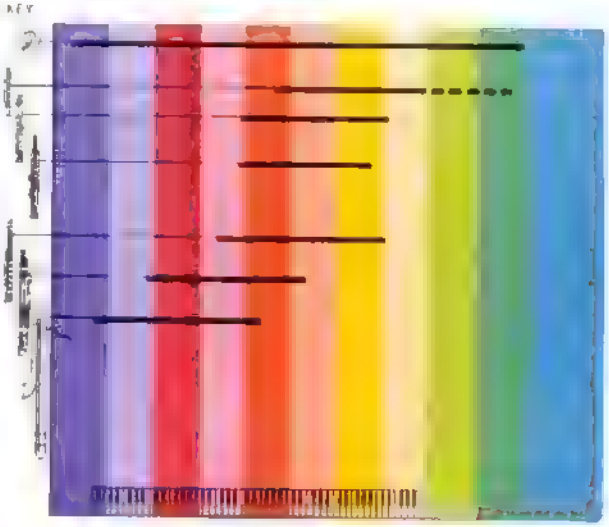
The nature of the vibrating object is the basis of family grouping of instruments. In string instruments – the violin, viola, cello, double bass, guitar, piano, harpsichord and harp – a taut string is vibrated by stroking it with a bow, plucking it with the fingers or a plectrum, or striking it with a soft hammer. A longer string produces a lower note, and the pitch of notes from an instrument is altered either by pressing the string against a fingerboard to change its length or by playing a string of a different length. The tension and thickness of the string also affect the note: a tauter or thinner string giving a higher note.

Types of musical instruments

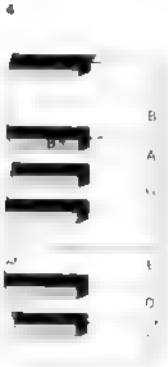
Wind instruments work by making a column of air vibrate. In brass instruments – the trumpet [5], trombone and horn – the player's lips vibrate in the mouthpiece. In some woodwind instruments such as the bassoon, oboe [2B] and clarinet [2C], the mouthpiece contains one or two vibrating reeds and in the flute [2A] the player blows across a hole to set the air column in the instrument vibrating. When a player presses down keys or valves, he alters the length of the air column and produces notes of different pitch [6]. Also, he can obtain some harmonics instead of the fundamental.

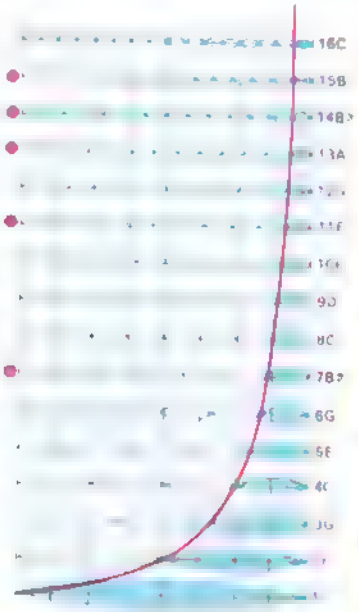
Some percussion instruments are played by striking either a taut skin, as in a drum, or a solid object of some kind – a disc of metal in a cymbal, for example. Tuned percussion instruments give definite pitches. They include the vibraphone [7] and xylophone, in which metal or wooden bars of different lengths are struck to sound various notes.

Electric instruments pick up the vibration of a string, as in an electric guitar, or a rod in an electric piano, and convert the vibration into an electric signal that passes to an amplifier and loudspeaker to produce the sound. Electronic instruments include the electronic organ and synthesizer, in which oscillator circuits produce electric signals.



Musical instruments, from the deepest to the highest members of each family, cover almost the entire range of human hearing. The woodwind family has a particularly wide compass, the lowest note of the contrabassoon and the harmonics (dotted line) of the piccolo nearing the limits of audibility.





4 Harmonics result from simultaneous vibration of smaller and smaller divisions of a vibrating string or air column. In the first 16 harmonics five (marked with dots) are out of tune.

5 A modern jazz-rock band contains all families of musical instruments, mingling the acoustic sounds of trumpet, saxophone and drums with the electronic sounds of synthesizer, electric piano, bass guitar and electric guitar.

6 Resonance is an important part of musical sounds. It can be demonstrated with a milk bottle and a tuning fork. The fork is struck and held over the neck of the bottle. The bottle can also sound, but its frequency depends on the length of the air column inside. It is possible to tune the bottle by adding a liquid until the remaining air space resonates at the same frequency as the tuning fork. The sound of the fork is then much louder. Many musical instruments make use of resonance. The low-volume sound produced by the vibrating string of a violin or guitar, for example, is made much louder by the air resonating inside the accurately shaped body of the instrument.

7 Sound production in musical instruments often involves resonance. Amplification of sound is achieved in the vibraphone in the same way as in the tuning fork and milk bottle. Beneath each bar is a tube of sufficient length to resonate at the frequency produced by the bar when it is struck. Small motor driven fans over the top of each tube blow air into it and at slow speeds produce a 'wavy' vibrato quality in the notes produced.

States of matter: gases

Everything in the world – all matter – exists in just one of three basic states. It is a gas, a liquid or a solid. Some substances can exist in all three states, depending on the temperature. Water, for example, is a liquid at ordinary temperatures. But above 100°C (212°F) it changes to a gas (steam) and below 0°C (32°F) it becomes a solid (ice).

All matter is composed of atoms or molecules and these are held together in liquids and solids by what are called intermolecular forces. The molecules are in continuous motion whose vigour depends on temperature. This "thermal motion" is restrained by the intermolecular forces of attraction, which hold the molecules together. Scientists call this view of matter the kinetic theory ("kinetic" means relating to movement or motion).

Kinetic theory of gases

In gases, thermal motion predominates and the molecules move rapidly in space, constantly colliding with each other and with the walls of their containing vessel. Collisions account for the pressure exerted by a gas [1].

Scientists have made measurements that confirm the kinetic theory of gases. At 0°C (32°F) and a pressure of 760mm (29.9in) of mercury, known as standard temperature and pressure (STP), a litre (1.75 pints) of oxygen is known to contain about 3×10^{22} (30,000 million million million) molecules. They move with a speed of about 430m/sec (1,411ft/sec). Molecules are extremely small and measured in ångström units (1 Å = 10^{-10}m). Each oxygen molecule is 3.5Å across. The molecules are, on average, 70Å apart, and they travel about 905Å between collisions; this distance is called the mean free path.

Boyle's law and Avogadro's principle

As temperature rises the kinetic energy of the molecules (their energy of motion) increases by an amount that is proportional to the change in absolute temperature. In a mixture of gases the average kinetic energy becomes the same for each kind of molecule.

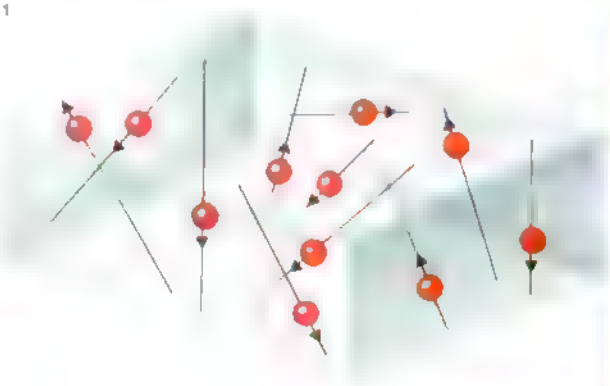
When a gas is compressed – that is, if its volume is made to change at a constant temperature – the volume is inversely proportional to the pressure. This relationship is known as Boyle's law [3] after its discoverer, the British scientist Robert Boyle (1627–91). It exists because, when the volume is reduced, collisions with the container walls become more frequent and the pressure rises. If the temperature of a gas rises but it is not allowed to expand (its volume is held constant), the pressure again increases because molecular collisions with the walls become more forceful as well as more frequent.

Another basic gas law is called Avogadro's principle after the Italian physicist Amadeo Avogadro (1776–1856). It states that, at the same temperature and pressure, equal volumes of all gases contain the same number of molecules. A litre of a dense gas such as carbon dioxide contains the same number of molecules as a litre of a light gas such as hydrogen.

Gases slowly diffuse through the walls of a porous vessel because their molecules are smaller than the minute holes in the walls of the container. The rate at which they do so is inversely proportional to the square root of

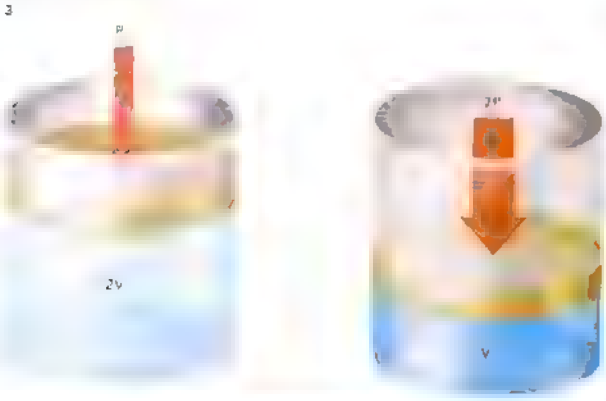
CONNECTIONS

See also

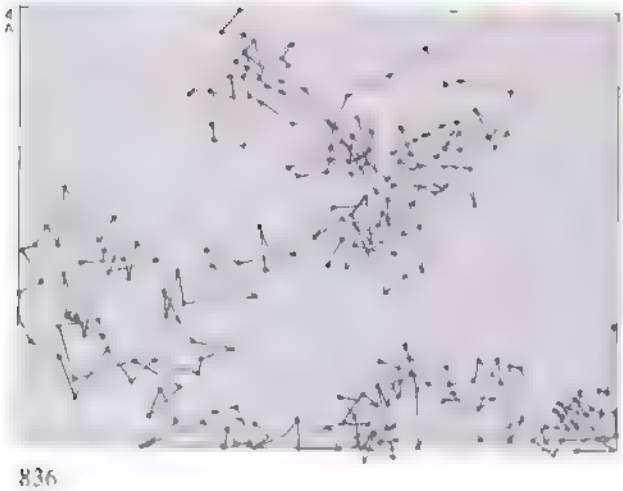


1 The molecules of a gas move continuously at various speeds and in various directions. Collisions with the walls of the container cause pressure and there are so many molecules in even the smallest volume that the pressure is the same everywhere in it. The actual pressure is proportional to the number of molecules in a unit volume and to the average kinetic energy (energy of motion) of the molecules.

2 A compressed gas is a store of potential energy and it can be made to do useful work when it expands to atmospheric pressure, as in pneumatic drills, hammers and "guns" for spraying paint.



3 When the pressure exerted by the piston is doubled, the gas volume is halved (provided temperature does not change). This is an example of Boyle's law: pressure is inversely proportional to volume.



4 In this diagram of Brownian movement [A] – which is evidence for kinetic theory – the dots represent the position of a particle recorded after equal small intervals of time and the lines joining them indicate the paths taken by the random motion of the particle. A beam of sunlight passing through smoke [B] is made visible by reflection from the smoke particles. The same principle is used to view Brownian movement with a microscope.

their density. Discovered by the British physicist Thomas Graham (1805–69), this relationship is called Graham's law [7]. It is explained by the kinetic theory: if different gases molecules have the same average kinetic energy, the lighter ones must move more quickly. Thus light gases diffuse through pores more quickly than dense ones.

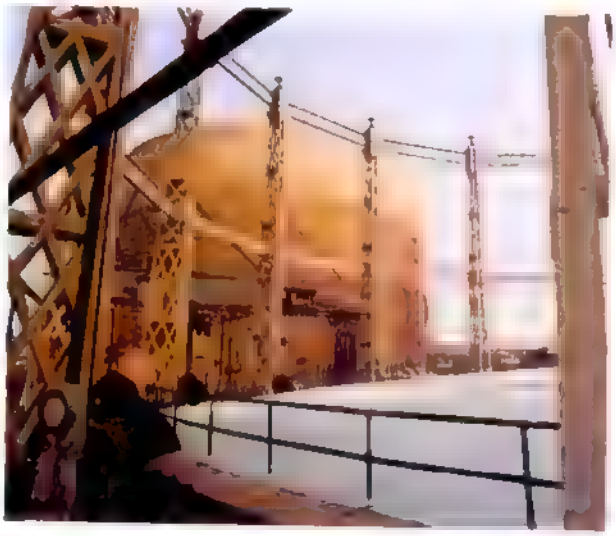
Brownian movement

More direct evidence for the kinetic theory is provided by the phenomenon known as Brownian movement [4]. Smoke can be seen in a sunbeam crossing a room. This effect can be produced in the laboratory by looking through a microscope at smoke particles in a box. Specks of light can be seen moving in a haphazard manner, first a short distance in one direction, then in another, and so on. The cause is unequal bombardment of the smoke particles, from different sides, by air molecules. The movement is less with larger smoke particles because they and the air molecules have the same kinetic energy, so the larger particles move more slowly.

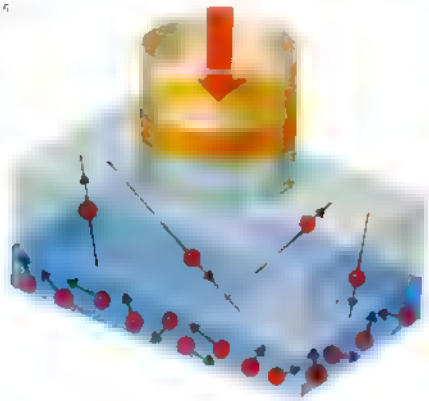
When a gas expands it has to do work against the external pressure. As a result it becomes cooler because the necessary energy must come from the kinetic energy of the gas. This phenomenon accounts for the coldness of the air escaping from a car tyre, for example. When such a change in pressure or volume occurs without heat entering or leaving the gas it is called an adiabatic change. The pressure changes in air as a sound wave passes are adiabatic.

The amount of heat needed to raise the temperature of unit mass of a substance through 1°C is called its specific heat capacity. A gas has two principal ones: that measured at constant pressure (c_p), and that at constant volume (c_v).

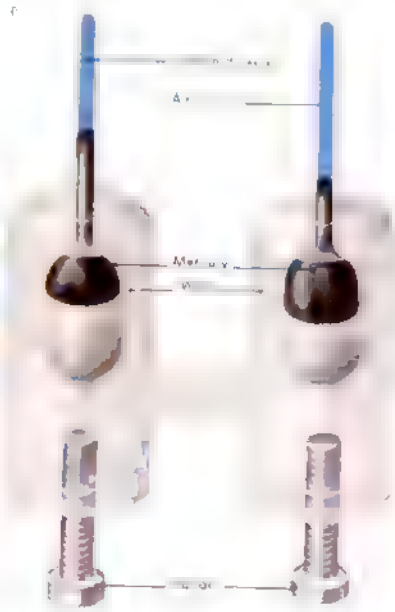
Subjected to sufficient pressure most gases turn into a liquid. But above a certain critical temperature it is impossible to liquefy a gas using pressure alone [5]. This is because, above this temperature, the kinetic energy of the molecules is sufficient to overcome the intermolecular attractions of their neighbours. Scientists therefore apply the cooling effects of adiabatic expansion in order to liquefy gases [6].



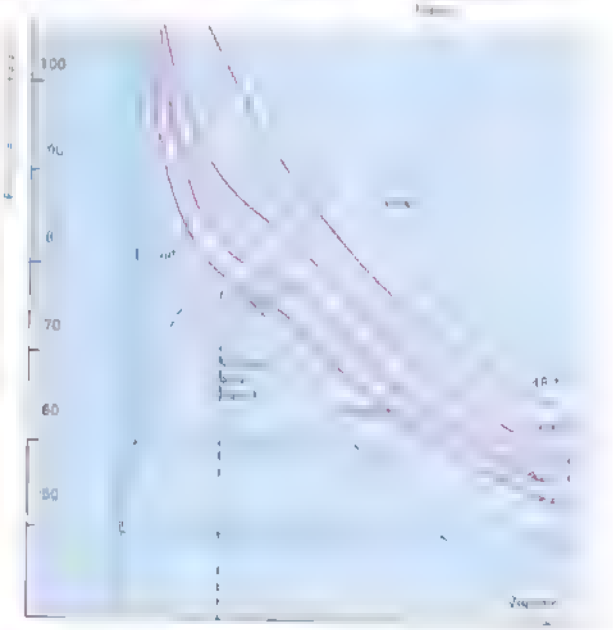
Gasometers store natural gas and do not [illegible] the same [illegible] measure quantity. They enable the pressure of the supply to consumers to be kept almost constant even with [illegible] demand. A gas [illegible] consists basically of a large movable cylinder with a water seal at the [illegible]

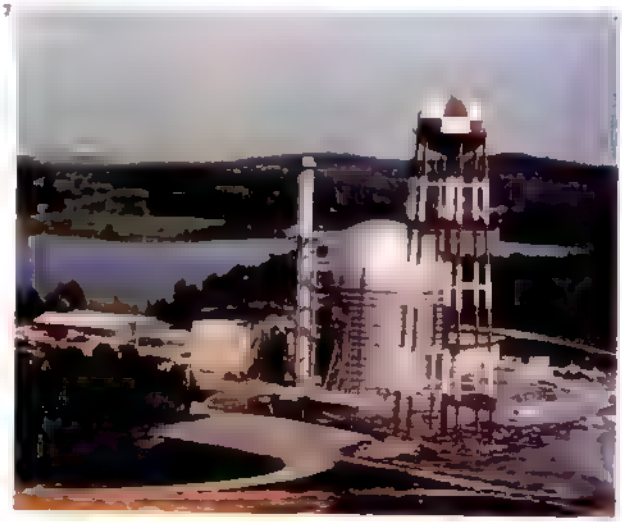


5 In early attempts to liquefy gases A, they were subjected to high pressure. Some gases liquefied under these conditions and a dynamic equilibrium was established between the liquid and vapour states – molecules left and entered the surface at the same rate. The experiments of Thomas Andrews (1813–85) B resulted in techniques that allowed other gases to be liquefied. Pressure is increased by screwing in the plungers and transmitted through the water to the gas and air in the upper tubes. The air is assumed to obey Boyle's law so that its change in volume is a measure of the pressure. Graphs of volume against pressure, called isothermals (C), reflect the various states of the gas. The horizontal blue line shows liquefication and does not appear until temperature falls below 31.1°C (86°F), the critical temperature of carbon dioxide.

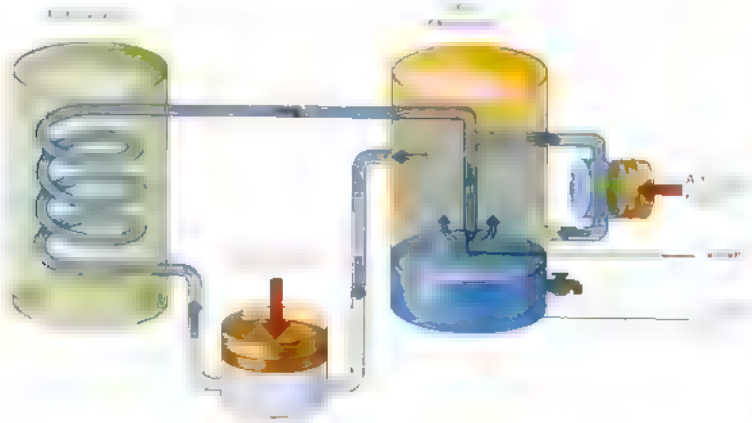


6 In an air liquefier, air free from water vapour and carbon dioxide is compressed and cooled by a refrigerator to −25°C (−13°F). It moves the piston part of an adiabatic engine and is further cooled to −160°C (−256°F). This air cools the other part of the high pressure air flowing down the centre pipes. Final cooling occurs by the Joule Thomson effect as the air expands through the valve.





7 The rate of diffusion of a gas is inversely proportional to its density. This principle (called Graham's law) was used to separate the isotopes of uranium during World War II to make the first atomic bombs and nuclear reactors in this plant at Oak Ridge in the USA. The uranium had to be converted into its fluoride, which is a volatile solid.



States of matter: liquids

A liquid occupies a definite volume and yet it can flow. The first property is evidence that a liquid's molecules are attracted to each other, whereas the second shows that they have greater freedom than those locked in the lattice of a solid. In a liquid, the molecules vibrate continually (at a rate of a million million times a second) and they change places with each other at nearly the same rate.

A stationary liquid cannot support any stress trying to shear it (as can a solid), because the pressure at any point is the same in all directions. The actual value of the pressure is the product of the depth, the density of the liquid and the acceleration due to gravity. For this reason, solid objects can float in a liquid and even a submerged one is acted on by an upthrust equal to the weight of the liquid displaced. This fact is known as Archimedes' principle.

Structure of liquids

Scientific methods used to study the structure of solids (such as X-ray diffraction) reveal that there are sometimes small volumes in a liquid with molecules in an ordered array. But there is no overall order as in a solid. In a hexagonal close-packed solid, for example, each molecule has 12 nearest neighbours. In a liquid the number varies between four and 11 and is continually changing.

The average distance between a liquid's molecules is greater than between those of a solid, which explains why most solids take up more room (expand) when they melt. But in a liquid the molecules cannot be squeezed close together (a liquid is almost incompressible). As a result, a liquid can transmit pressure along a pipe [5].

Evaporation and boiling

When a liquid is heated, its molecules move more and more until, at the boiling-point, the liquid turns into a gas or vapour. The heat energy needed to vaporize it is called the latent heat of vaporization. Similarly, when a liquid is cooled, its molecules move less quickly until they take up fixed positions and the liquid freezes into a solid. The heat needed to melt a frozen solid back to a liquid at the same temperature is called the latent heat of fusion.

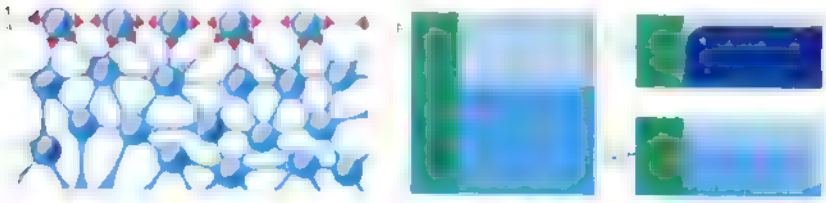
Even at ordinary temperatures (below boiling-point) some molecules "jump" out of the surface of a liquid to form vapour – they evaporate. In a closed vessel there is an equilibrium between a liquid and its saturated vapour; the rate at which molecules leave the liquid is the same as that at which they re-enter it.

When a liquid boils, some work has to be done by the escaping vapour to overcome atmospheric pressure. If the pressure on a liquid is reduced it boils at a lower temperature. If the pressure is increased the boiling-point rises. But if there are no tiny particles in the liquid on which vapour can form bubbles, boiling is suppressed. This effect causes the "bumping" that takes place when pure water is boiled and is the principle of the liquid hydrogen bubble chamber [6]. Water does not follow many of the general rules that apply to liquids [3]. Most substances expand by between five and 15 per cent on melting, for instance, but water contracts by about ten per cent and expands on freezing. These properties arise from the highly directional nature of intermolecular forces in water (due to hyd-

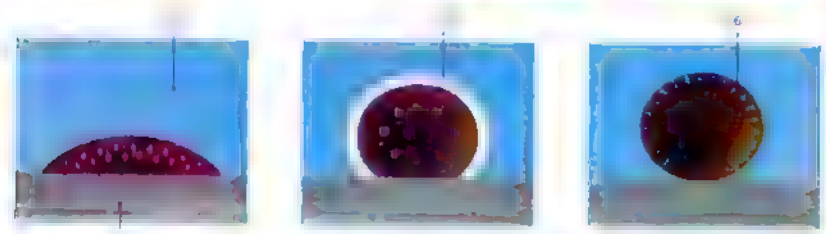
CONNECTIONS

See also

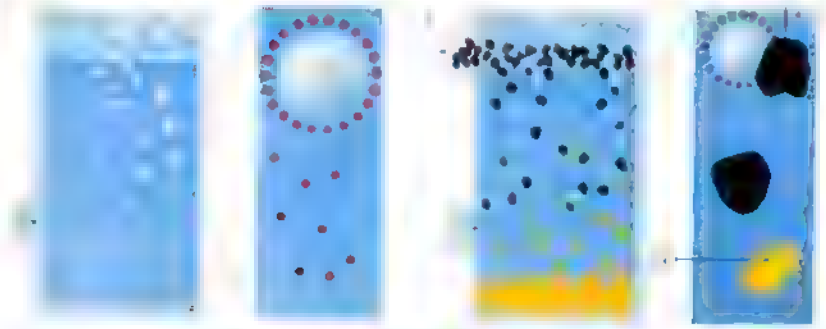
States of matter
Pressure and flow
Heat and
Liquids and gases



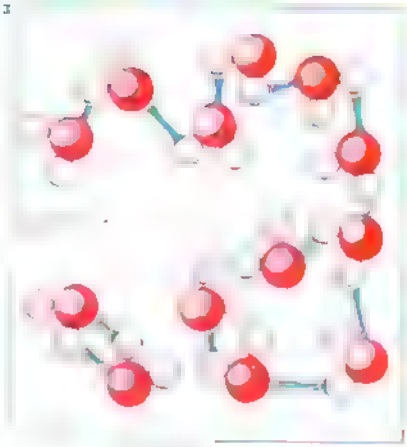
1 At the surface [A], the force between a liquid's molecules causes the surface to behave like a stretched membrane. The surface in a glass container [B] curves depending on whether the glass attracts liquid weakly (eg mercury [1]) or strongly (eg water [2]). An oily liquid [C] adhering to cloth [3] is weakly attracted by water [4], a detergent forms a new surface layer [5] attracted to both (detergent [6] is partly oil-like and partly salt-like). Some detergents [D] prevent gas bubbles [7] collapsing, making it possible [E] to separate ore [8] clinging to them from earth and sand [9] by flotation.



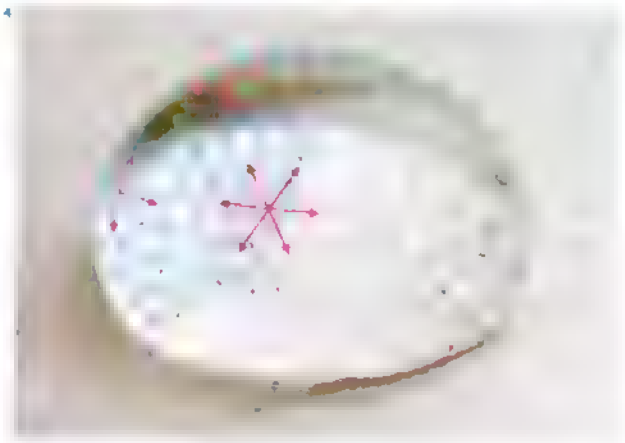
2 Oil is a prime source of energy and chemicals, and huge quantities of crude oil are carried in tankers from the producing countries to highly industrialized ones, such as the USA, Japan and Western Europe. By Archimedes' principle a tanker displaces its own weight of seawater in order to float and the cargo (oil) is less dense than water. The tanker sinks slightly deeper in warmer water and on the return journey water ballast is carried in the tanks to keep the vessel stable. If a mishap occurs, escaping oil floats on the water; oil slicks are a form of pollution.



3 In all materials the atoms or molecules are in continuous motion. The energy of this motion depends on the temperature. In a liquid, the motion prevents any permanent intermolecular structure from forming, but forces of attraction govern the overall volume. In water there are many temporary linkages (shown blue) between the molecules; very small cavities [1] form and vanish, giving water an ever-changing structure.



4 All the molecules in a liquid exert attractive forces on their immediate neighbours. Within the main body of the liquid the effects of these attractions cancel each other out. But on the surface the attraction can take place only inwards (there are no molecules outside the surface to counteract it). As a result the surface is in tension. This surface tension behaves like a "skin" which pulls a droplet of water into a spherical shape.



rogen bonding). In the solid state (ice) it produces a very open structure, which disappears on melting. Water has its maximum density at about 4°C (39°F), probably due to a crystal-like ordering of small groups of molecules which disappears on warming.

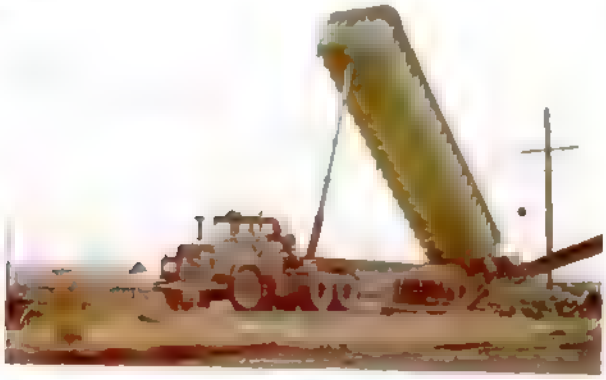
Surface tension and viscosity

In the centre of a liquid each molecule is attracted by all those surrounding it and their net effect is zero. But at the surface there can be little upward force to balance the attractive downward forces. As a result a surface molecule tends to be pulled into the body of the liquid [4]. The number of molecules at the surface becomes the smallest possible, and the surface behaves as if it were in tension and had a "skin" on it. The membranous effect of surface tension allows small, dense objects, such as needles or insects, to "float" on the surface of water. If the cohesive forces of attraction between a liquid's molecules are large, it has a high surface tension and a large viscosity (stickiness).

Water wets glass because the cohesive force between a water molecule and a glass molecule is greater than that between two water molecules [1]. The opposite is true of glass and a liquid such as mercury. A liquid such as water rises in a fine capillary tube dipped into it and the meniscus (shape of the surface) curves inwards, or is concave. Mercury, on the other hand, is depressed in a capillary tube and possesses a convex (outward curving) meniscus.

Water pours more easily from a jar than does treacle and treacle is said to be more viscous than water. A simple model of what happens shows one layer of liquid molecules under a shearing stress sliding over another layer. For this to happen any molecule in the faster-moving layer must overcome the attraction of the nearest one in the adjoining layer. And having moved one place along the line it must repeat the process. To do this it must use some energy and this slows it down. The relative velocity between the layers is reduced and the result is viscosity. Heating a liquid provides it with more energy and, as expected, viscosity falls with increasing temperature. For this reason, warm treacle pours more easily than cold treacle.

Various liquids differ in their physical properties such as boiling point and viscosity (stickiness). Water and wine have roughly the same boiling point and low viscosity. Oil is more viscous and boils at a higher temperature whereas exceptionally viscous liquids such as honey and tomato ketchup have extremely high boiling temperatures.



5 Liquids are almost incompressible and as a result can transmit pressure. This important principle finds many applications in the branch of engineering called hydraulics. Many trucks have an hydraulic jack in which pressure transmitted by means of oil is used to lift the load. A pump is used to provide the pressure and provision has to be made for the oil to run back when the pressure is released.

6 In a bubble chamber, a dust-free liquid in a perfectly clean vessel is heated to a temperature above its boiling point and extra pressure is applied to stabilize it. If charged particles are then directed into the chamber, bubbles form on the charged "nuclei" left by the particle along its track, making it visible. Liquid hydrogen is generally used as it is a good source of protons on which bubbles can form.

7 A water drop at the end of a glass tube takes its shape because of surface tension. The attraction between its molecules leads to a spherical shape as the "skin" effect caused by inward-acting forces in the surface holds the bulk of the liquid back. But a water droplet becomes a distorted sphere because gravitational as well as surface tension forces are acting on it.

8 The most common method of measuring temperature makes use of the expansion of a liquid on heating. A mercury thermometer (left) has a wide range (−39°C to 360°C), but needs a large bulb (reservoir) and a narrow stem if small temperature changes are to be detected. Alcohol can be used in thermometers for measuring lower temperatures; it expands more but boils at 78°C. The great advantage of liquid-in-glass thermometers is that they can be read directly and carried easily.

9 Motor oil is more viscous than water, as can be seen when each is poured; the water flows much more easily than oil because layers of water molecules slide over each other more easily than do layers of oil molecules. Normally, the viscosity of a liquid decreases with rise in temperature. Oil for a particular application must have the right lubricating properties and much research has been done to produce the correct oils for car engines and gearboxes, including oils whose viscosities change only a little when they get hot. When a liquid flows in a pipe in streamlined motion the region in contact with the pipe is still and that near the axis has the greatest velocity.

States of matter: solids

The crystalline shapes of many solids indicate that the atoms in them take up some kind of regular arrangement. In the amorphous or non-crystalline substance, there is no regular order. There are seven main crystalline structures, of which the cubic system is the simplest. Sodium chloride (common salt) is composed of sodium ions and chloride ions. In the solid salt these ions take up what is called a face-centred cubic structure. This and other arrangements can be confirmed by making the crystals diffract an X-ray beam, then experts can use such X-ray photographs to work out the structures of complex crystals.

Sodium chloride is an example of an ionic crystalline substance. Other crystalline substances, such as diamond, consist of a regular array of atoms linked to each other by covalent chemical bonds, in which one or more electrons are shared between neighbouring atoms. In waxes and similar substances, molecules are held together only weakly by what are called Van der Waals forces. And a metal has a lattice of positive ions in which free electrons occupy the spaces. Applying a voltage across the metal makes electrons drift between the ions, which is the reason why metals and their alloys are good conductors of heat and electricity.

Vibrating and slipping atoms

All intermolecular forces can be thought of as electrical in origin, causing attraction between molecules at relatively large distances and causing repulsion at close quarters. The elastic properties of solids can be explained in terms of such forces. When a material is stretched, the distances between its atoms increase slightly and the resulting strain is found to be proportional to the stress producing it (the relationship known as Hooke's law). Compression moves the atoms closer together, while shearing makes layers of atoms slide over each other.

The atoms in a solid – even a crystalline one – vibrate about their average position in the lattice. Heating a pure solid makes its atoms vibrate more vigorously. Sufficient heat energy overcomes the forces holding the atoms together, the crystalline structure "falls apart" and the solid melts.

A single crystal of a pure metal is much weaker than might be expected. This may be due to imperfections in the lattice which cause dislocations [7]. Under stress, the layers of atoms move in such a way that the dislocation shifts towards the edge. An ordinary metal is polycrystalline, consisting of an irregular arrangement of many small crystals. Stress makes layers of atoms in individual crystals slip over each other. But atoms of an impurity within a crystal can "anchor" dislocations and prevent slipping. Consequently an alloy is usually harder than the two or more metals of which it is composed.

Metal fatigue: cause and cure

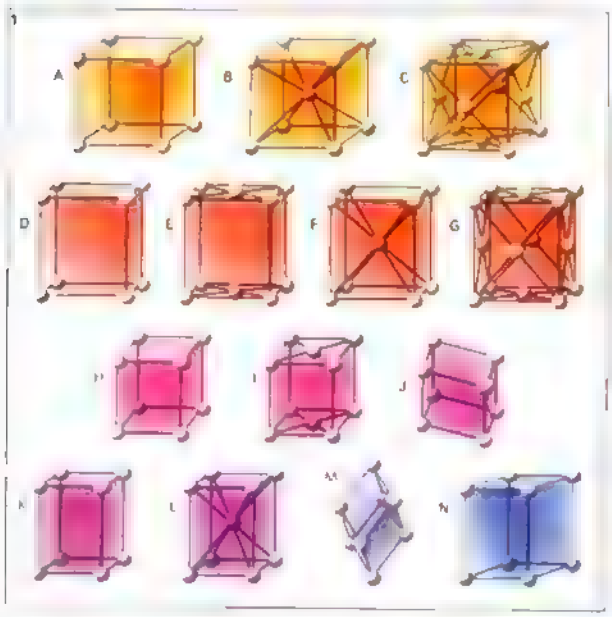
Strain deforms some solids over a period of time in the phenomenon called creep. This can be due to the movement of dislocations in crystalline grains, slip between grain boundaries, or slip along well-defined glide planes.

"Fatigue" is the name given to a change in metallic properties that may result in sudden breakage. It resembles work-hardening (caused by hammering) in that dislocations in crystallites interlock, causing brittleness.

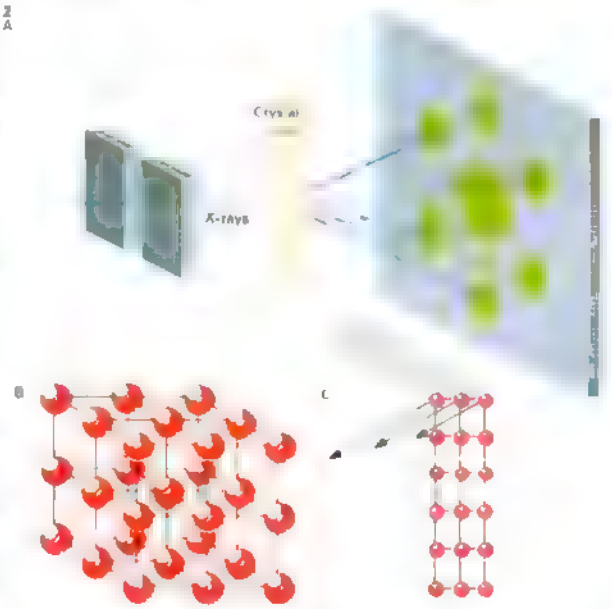
Scientists can make thin fibres of sub-

CONNECTIONS

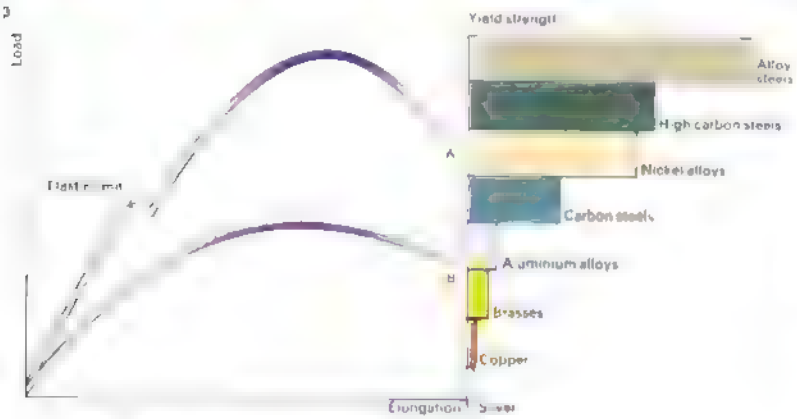
See also



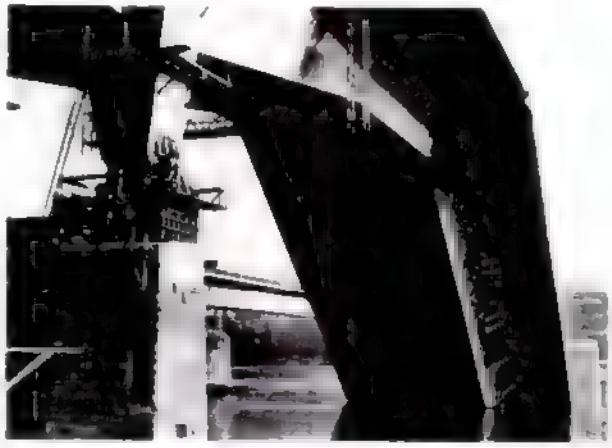
1 Crystals are formed by the regular stacking of identical building blocks. The elementary block is called a unit cell and may contain a single atom such as in copper, or it may contain hundreds or thousands of many different types, as in some protein crystals. The regularities and symmetries in crystals are a great help in identifying and solving their structure. The unit cells of all crystals can be classified into one of 14 basic types called space lattices. They are conveniently grouped into seven systems: cubic (A, B and C) orthorhombic (D, E, F and G), monoclinic (H and I), triclinic (J), tetragonal (K and L), trigonal (M) and hexagonal (N).



2 X-rays diffract from the periodic regular array of unit cells in a crystal to produce a pattern of diffraction spots (A). X-ray waves reflecting from successive planes need to be a whole number of wavelengths out of step in order to produce a diffraction spot (C). Symmetries and regularities in the pattern help the crystallographer to decide which type of unit cell is present. For instance the diffraction pattern in (A) shows hexagonal symmetry, so a unit cell such as those in (B) cannot be present. From the relative intensities of the spots and their phases one can calculate the crystal structure.



3 Tensile strength is tested by stretching. A metal extends overall at first, but later expansion concentrates around the point of fracture. Curves A and B show typical extensions. Curve A for mild steel remains linear to its elastic limit. If the load is released early, the metal returns to its original length. Curve B is typical of softer metals. A range of relative strengths of metals is also shown (right).



4 Engineers use high grade steel girders in their work because good steel is usually highly resistant to cracking. If a crack does appear, the metal is generally ductile enough for the edges of the crack to flow together, which will diminish the danger of the crack extending far. In poor steels cracks may develop rapidly. This is a particular danger in bridges where the steel has to resist changes in temperature.

stances free from any dislocations. Called whiskers, they are very strong and when incorporated into a matrix of another substance produce a strong composite material.

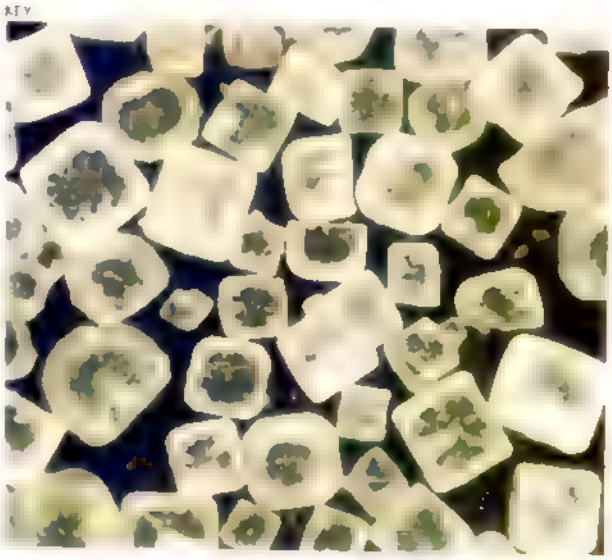
Structural effects in solids

Molecules in natural and man-made polymers have a complex arrangement. X-ray photographs reveal that, when rubber is stretched, its coiled long-chain molecules line up. When tension is released, the molecules snap back into their former shapes, and as a result giving rubber its elasticity. Similarly a stretching process during the manufacture of nylon lines up long chains of molecules.

Semiconductors – the key to modern solid-state devices such as transistors – make use of subtle variations in an otherwise normal crystal lattice. The basic element, such as silicon or germanium, has incorporated in it minute traces of a deliberately introduced impurity element. The impurity has either one more or one fewer electron in its atoms than the basic semiconductor element. As a result, there is a slight excess of electrons or a slight deficiency (an absence of an electron is in this context called a "hole"), and it is the movement of these electrons or holes that gives the materials their special electrical properties. When extra electrons are present the semiconductor has negatively charged current carriers and so is called an *n*-type semiconductor. The holes in the second type are considered to be positive carriers and it is called *p*-type.

X-ray diffraction enables scientists to study the microstructure of solids. Other methods reveal the overall or macrostructure. The grain structure of a metal, for example, can be revealed by etching the surface and viewing it by reflected light through an optical microscope.

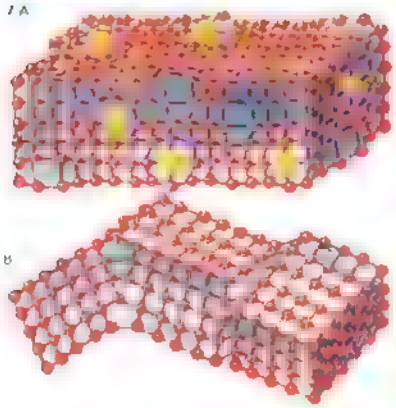
Higher magnifications are possible using an electron microscope. Scientists make a copy of the etched surface by depositing on it a layer of carbon or plastic and stripping it off as a thin film to be viewed with the microscope. They may use a very thin foil of the metal to be studied. The fine structure of the surface is revealed in three-dimensional detail by a scanning electron microscope.



Seen through a microscope, salt crystals are revealed as regular cubes. Perfect – that is unbroken – crystals have precise shapes (although here tumbling of the crystals has knocked the corners off some of them). There are six other basic shapes.

5 Metalworkers hammer sheet metals to harden them without making them brittle. The hammering moves dislocations along intersecting slip planes until they meet and stop. These meeting places act as barriers to the movement of any other dislocations, making the metal stronger.

6 Foundryworkers separate metals from ores by smelting them. Metals are heated to break down the lattice of atoms so that the metal flows. They are poured into moulds or cooled and rolled into sheets. Melting points of metals range from mercury which melts at −38.8°C (−38°F) to tungsten which melts at 3,410°C (6,170°F).



7 Strong metals restrict the free movement of dislocations. A metallurgist may achieve a strong metal by making an alloy, or he may make the metal's crystals as small as possible. In A large atoms are at the crystal corners and small ones in the lattice centre, distorting the whole crystal and preventing any free dislocation movement. In B the crystal boundaries are mismatched and they too have the effect of creating barriers against dislocation movement. Where dislocation of the atomic lattice occurs, or where there is slipping between grain boundaries and along glide planes the metal will fail. Frequently repeated strains and fluctuating loads may eventually cause metal fatigue (C). The edges of the fractures may show signs of metallurgical recrystallization. Strain over a long period produces a similar effect, called creep failure (D). The example of failures C and D occurred in a nickel alloy turbine blade.

8 Increased pressure lowers the melting-point in substances such as water, which expand when they solidify. Ice melting under the pressure of a skate acts as a lubricant that makes the skater's motion both smooth and easy.

Heat and temperature

A bicycle pump becomes warm when pumping up a tyre, an effect that can be explained by the kinetic theory of heat, first suspected by Isaac Newton (1642–1727) and developed by eighteenth and nineteenth century scientists. The theory describes heat as the kinetic energy (energy of motion) of the vibrating atoms or molecules that make up every substance. In the bicycle pump, air molecules are speeded up by collisions with the pump's piston. The increase in their kinetic energy takes the form of heat.

Thermal agitation and molecular motion

According to the third law of thermodynamics, absolute zero – the temperature (−273·16°C) at which molecules cease all motion – is unobtainable. As a result, molecules perform a continuous motion known as "thermal agitation", which increases in vigour as heat is transferred to an object. Indirect evidence of this incessant motion was first obtained by the botanist Robert Brown (1773–1858) in 1827. He discovered that tiny pollen grains suspended in water were continually making jerky movements. It is the continual unequal bombardment of each tiny speck by the molecules of the liquid that produces this "Brownian motion". The smaller the particle, the more violent is the motion.

The kinetic theory also explains why, when a hot gas is mixed with a cooler one, a common temperature is eventually reached. Molecular kinetic energy of the hot gas molecules is transferred by thousands of collisions to the cold gas molecules, until the average kinetic energies of both gases are the same. The molecules all travel with different velocities, changing after each collision. For this reason the temperature of the gas (or other substance) is a measure of the average molecular kinetic energy.

The changes of state that may occur when a substance is heated can be described by the kinetic theory [1]. In a solid, the atoms or molecules are tightly bound together and vibrate only about an average position. As the solid is heated its internal kinetic energy increases and the particles vibrate more vigorously. They move farther and farther apart until eventually the attraction between them is insufficient to keep them in a fixed position. They can then slide about and exchange partners – the solid melts and a liquid is formed. The heat needed to achieve this change of state from tightly bound solid to loosely bound liquid at the same temperature is called the latent heat of fusion. If yet more heat is applied, the atoms or molecules gain more kinetic energy, move with greater velocity within the liquid and the proportion of these escaping from the surface increases (vapour pressure increases). Eventually, at the boiling-point, so many atoms have enough energy to escape from the liquid that the vapour pressure equals atmospheric pressure. In the gaseous state the atoms or molecules move almost independently, the conversion of a liquid at its boiling-point into a gas requires latent heat of vaporization.

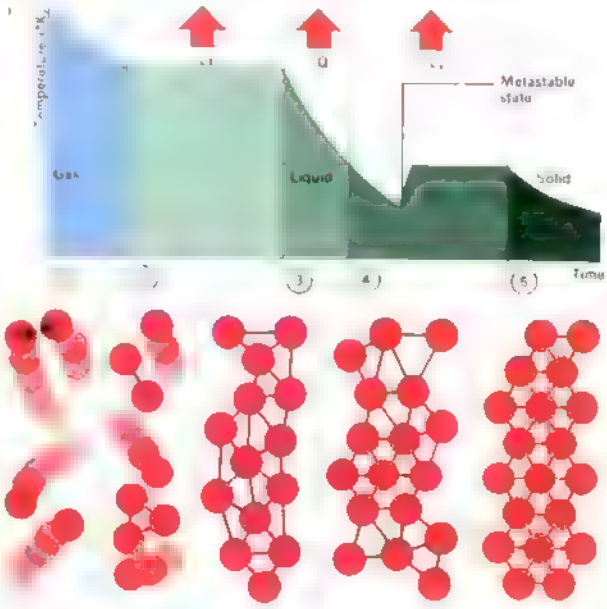
Changes of temperature

Instead of changing the state of a substance (for which latent heat is needed), applied heat energy may merely raise its temperature. The temperature change depends directly on the quantity of heat transferred

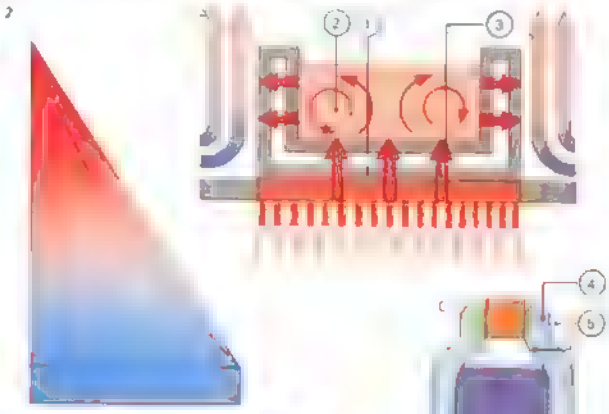
CONNECTIONS

See also

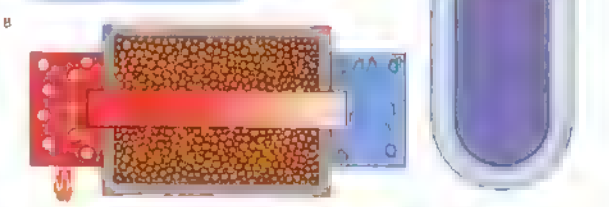
Stir Tyndall

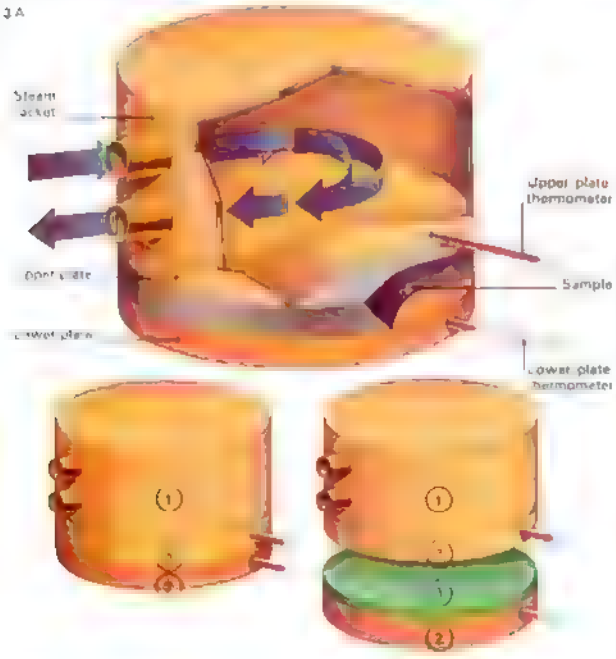


1 Vibration of atoms and molecules in a substance governs its temperature. In a gas [1] atoms move independently and their average velocity and mass determine the internal energy and temperature. After cooling loss of plant heat of condensation [L_v] converts the gas at boiling-point to a liquid state [2] when its atoms become locked in a weakly bonded arrangement. Further cooling [3] to freezing-point loses a quantity of heat [Q] and a solid then forms through [4] release of latent heat of fusion [L_f]; the atoms then become rigidly bonded together [5]. A colder "metastable" phase can precede actual freezing.

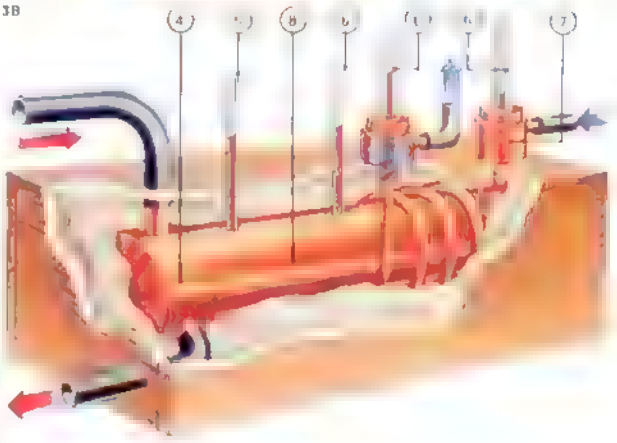


2 The three ways in which heat moves all take place when a pan is heated [A] – conduction through the metal walls of the pan [1], convection by fluid motion [2] and radiation from the heat source to the pan [3]. In theory an insulated good conductor with ice at one end and boiling water at the other varies in temperature linearly with distance along the bar [B], as in the straight-line graph. With poor insulation a curve like the dotted line results. A vacuum flask [C] has a vacuum [4] to prevent conduction and convection and silvered walls [5], to minimize heat loss by radiation.





3 Conductivity is the amount of heat passing in unit time across unit cross-section per unit of temperature gradient. It can be measured [A] by noting the time taken for a known quantity of heat to pass through a sample. Two plates of equal area [2] are put against the material [3]. Thermometers on each side measure the temperatures as the upper plate is heated by a steam jacket [1]. Materials of high conductivity [B] are tested in a cylinder [8] heated at one end by steam [4] to 100°C [5]. Other thermometers [6] measure the temperature of the sample and the rise in temperature of water circulating through a jacket attached to the other end [7], and give a way of calculating conductivity.



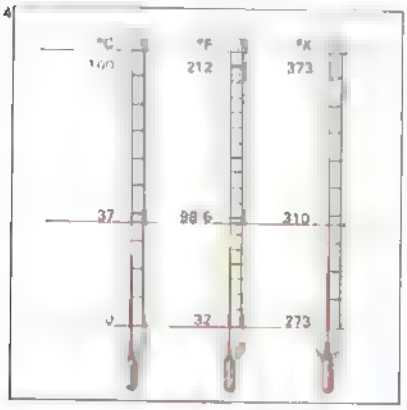
and this is measured in units called calories or joules (4.2 joules=1 calorie). The calorie is defined as the quantity of heat that raises the temperature of 1 gramme of water by 1°C. So the quantity of heat needed to raise the temperature of 1 gramme of any substance by 1°C can be measured in these units – and this is called the specific heat of the substance. The quantity of heat that raises the whole bulk of substance by 1°C is called the thermal capacity of the given mass.

It is possible for heat to be transferred from place to place and there are three ways in which this can occur: by conduction, convection and radiation [2]. The first two methods rely on the fact that atoms that have received kinetic energy from a heat source can, in collision, transmit this to their neighbours. In a tightly bound solid only nearest-neighbour collisions occur and any heat transfer through the solid is called conduction. In a fluid (liquid or gas), the medium itself can move and transport atoms of high kinetic energy to the cooler parts of the fluid where they then transfer their heat – this is convection. Even when no physical contact exists between atoms, heat can still be transferred. For instance, heat from the Sun reaches the Earth through the near vacuum of space. This method is radiation. Different substances conduct heat at different rates. Their ability to transmit heat is known as their thermal conductivity [3].

Measuring temperature

All methods of measuring temperature changes are based on the ways in which materials change physically when heated. The most commonly used characteristic is the expansion of solids or liquids when heated. Usually the physical change caused by a temperature change is made visible against a calibrated scale on the measuring instrument or thermometer. In making any type of thermometer [5, 6, 7], two constant temperatures or "fixed points" must first be marked. The range between them can then be subdivided as finely as desired. The numbers assigned to the fixed points and the numbers of degrees between them define the temperature scale, such as Celsius (or centigrade), Fahrenheit and Kelvin (or absolute) [4].

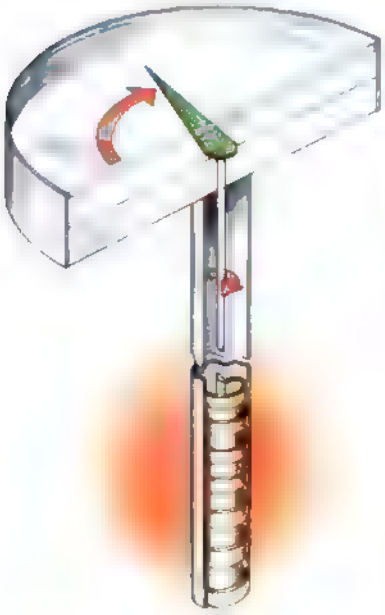
A furnace converts the chemical energy in coal to heat energy used to produce the steam that drives an engine. The heat of the furnace brings the water to its boiling-point and provides latent heat of vaporization to turn it into steam.

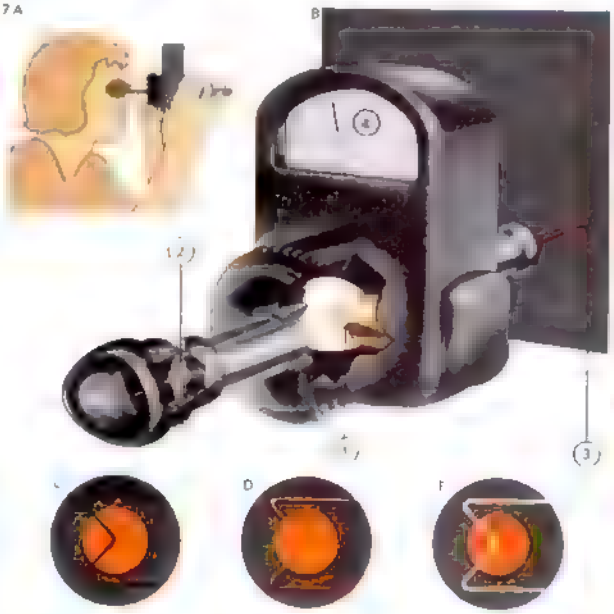


4 Temperature scales are arbitrary both in their range and in their division into degrees. The freezing-point of water is set at 32° Fahrenheit (°F), 0° Celsius or centigrade (°C) and 273° Kelvin (°K). Between this point and the boiling-point of water there are 180°F, 100°C and 100°K. The Kelvin scale is unique in setting its lowest fixed point at absolute zero, which in practice is unobtainable.

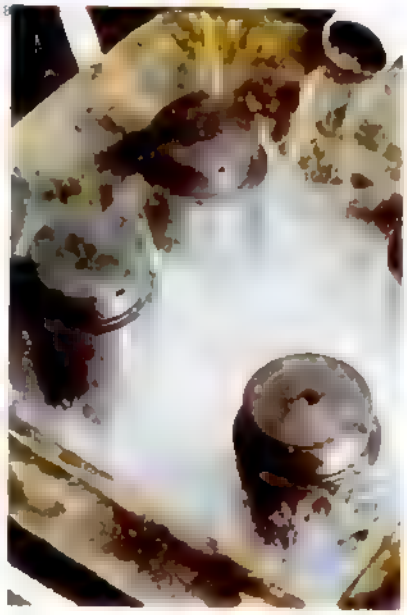
5 The indicators in a maximum and minimum thermometer (A) are pushed along by mercury in a U-tube and stay put at the farthest point of travel. Liquid in a cylindrical reservoir senses the temperature and its contraction (B) or expansion (C) displaces the mercury along scales at each side of the U. The steel indicators can be reset simply by moving a small magnet along the outside of the tube.

6 A bimetallic strip thermometer uses a helical metal strip that unwinds when heated and rotates a pointer over a calibrated scale. When warmed, the inner metal (usually copper) expands more than the outer one, so causing the bimetallic strip to unwind. The material used for the outer strip is usually Invar, an alloy of iron and nickel which has a low coefficient of thermal expansion.





7 An optical pyrometer allows very high temperatures to be measured at a distance from the temperature source (A). It exploits the fact that two solids at the same high temperature radiate light with the same spectrum and show the same colour. To measure the heat of a furnace (B), an electrically heated filament (1) fixed in the tube of a telescope with a special lens (2) is heated until it glows with the same colour as that emitted by the furnace (3). Comparison with the background image shows if the furnace is hotter (C), as hot (D) or colder (E) than the wire. A meter (4) indicates the current passing through the wire and this can be calibrated directly in temperature degrees.



8 Liquid air can be produced by cooling air, which is normally gaseous at room temperature. During the change to a liquid state, latent heat of condensation is released. To cut down heat input the liquid is kept in a Dewar vacuum flask. But if it is poured out of this flask, heat from the surroundings makes the liquid boil rapidly. Large amounts of gas are generated during this process and cause a fog of condensed water vapour to form above the liquid.

Order and disorder: thermodynamics

Thermodynamics (meaning "the movement of heat") deals with the ways in which heat energy travels from one place to another and how heat is converted into other forms of energy. In a heat-transfer process temperature, pressure and volume may each or all undergo various changes. Much of thermodynamics consists of ways of mathematically manipulating these and other parameters to be able to make predictions about the ways in which they will and do change.

The four laws of thermodynamics

Historically scientists first derived three laws called the first, second and third laws of thermodynamics. Then an even more fundamental law was recognized. It has been labelled the "zeroth" law of thermodynamics.

If a hot and a cold object are brought into contact, they finally reach the same temperature [1]. The hot object emits more heat energy than it receives and the cold object has a net absorption of heat. Both objects absorb and emit energy continually, although in unequal quantities, and the exchange process continues until the temperatures equalize. Each object is then absorbing and emitting equal amounts of heat and the objects are said to be in "thermal equilibrium". The zeroth law states that, if two objects are each in thermal equilibrium with a third object, then they are in thermal equilibrium with each other.

The first law really has two parts – the first is the law of conservation of energy and the second effectively defines "heat energy" and how types of energy can be converted into one another [2]. If heat energy is supplied to a system, then the first law states that this equals the change of internal energy of the system together with the mechanical energy that allows the system to do external work. Thus in a petrol engine an air/petrol vapour mixture is ignited after being compressed. The burning of the mixture releases heat energy from the chemical reaction, thereby causing the gases to expand and do work against the piston by moving it. The burned gases are finally hotter than the system was before the explosion, so there is a change in the internal energy of the petrol engine system. The sum of this energy change and the external work done equals the released heat energy.

Following the zeroth law, which defines temperature, and the first law, which describes energy conversion, the second law of thermodynamics governs the direction of flow of heat energy between objects at different temperatures. It says that, of its own accord, heat can flow only from a hot to a cold object. The heat transfer increases the motion of the molecules of which the colder object is composed and so effectively increases its internal "disorder".

Cooling substances

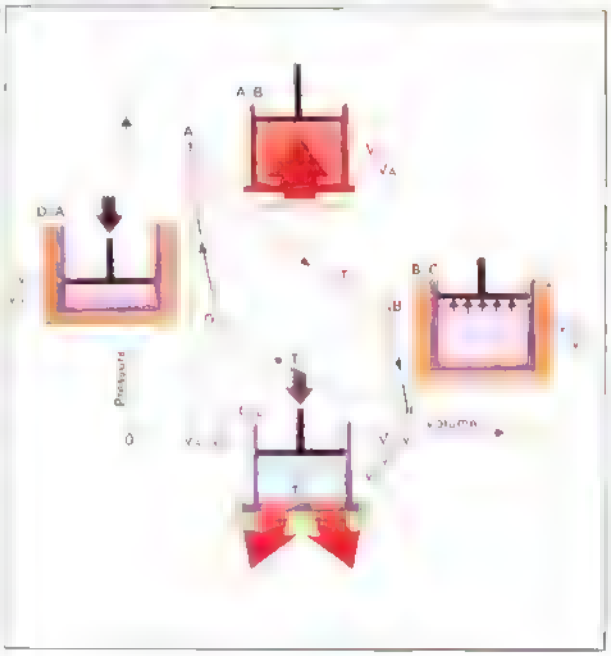
The third law of thermodynamics states that it is impossible to cool any substance to absolute zero. This zero of temperature would occur for example in a gas whose pressure was zero. All its molecules would have stopped moving and possess zero energy, so that extracting further energy and achieving corresponding cooling would be impossible. A substance becomes progressively more difficult to cool as its temperature approaches absolute zero (−273.16°C).

CONNECTIONS

See also

1 The thermic lance produces enormously high temperatures. If it is directed on to a substance the intense heat rapidly melts the substance in the region of the lance so that the substance becomes fluid and can be carved at will as the lance advances. Here a lance is being used to cut through a huge concrete block.

2 Friction occurs whenever two surfaces rub together and this in turn generates heat. The sudden surge of power at the beginning of a dragster race spins a car's rear wheels and the heat generated burns the rubber of the tyres. Similar heating occurs in a car's brakes when they are used.



3 Carbon dioxide, released as a solid from a fire extinguisher, turns into a gaseous form and prevents oxygen reaching the fire.

4 Heat can be converted to mechanical work by allowing a heated gas to expand. Maximum efficiency comes from the Carnot cycle of alternate adiabatic [BC DA] and isothermal [AB CD] processes. The former take place without gaining or losing heat, the latter without changing temperature. From A to C the gas raises the piston and performs work equal to the area below the curve, that is V_aABCV. From C to A it uses work equal to area V_aADCV_c. The net work obtained equals area ABCD.

From the statement of the second law, a heat transfer process naturally proceeds "downhill" – from a hotter to a cooler object. There must be some property or parameter of the system that is a measure of its internal state (its order or disorder), and which has different values at the start and the end of a possible process (one allowed by the first law). This parameter is termed "entropy", and the second law maintains that the entropy of an isolated system can only remain constant or increase.

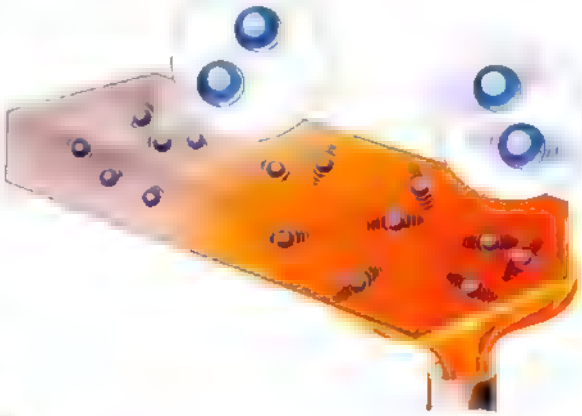
Careful observation of machines shows that they consume more energy than they convert to useful work. Even if no energy is wasted in friction or lost by necessity, as in a radiator, the available mechanical energy is less than that supplied by the heat source. The entropy of the system is a reflection of its inaccessible energy, and the second law says that it cannot decrease. Heat is a random motion of atoms and when the energy is degraded towards the inaccessible energy pool, these atoms assume a more disorderly state – and entropy is a measure of this disorder.

Under the constraints imposed by the laws of thermodynamics it is possible for a system to undergo a series of changes of its state (in terms of its pressure, volume and temperature). In some cases the series ends with a return to the initial state, useful work having been done during the series.

Heat cycles and efficiency

The sequence of changes of the system is called a heat cycle and the theoretical maximum efficiency for such a "heat engine" would be obtained from following the so-called Carnot cycle [4] which is named after the Frenchman Nicholas Carnot (1796–1832). If it were possible to construct a machine operating in cycles which, in any number of complete cycles, would generate more energy in the form of work than was supplied to it in the form of heat then the dream of the perpetual motion machine would be possible [5]. The first law states the impossibility of achieving this result and the second law denies the possibility of even merely converting all the heat to an exactly equivalent amount of mechanical work.

KEY
A flame applied to one end of a metal bar transfers heat energy to the atoms of the metal. This raises their kinetic energy so that the atoms begin to vibrate increasingly more vigorously about their fixed mean positions within the lattice network of the metal. As the extent of vibration increases, collisions with neighbouring atoms occur so that energy is transferred to these atoms causing them to vibrate. Heat energy is eventually transmitted to the other end of the bar and if the flame is kept in position for some time the temperatures at each end of the bar will tend to equalize.



5

5 Perpetual motion machines can be classified according to which of the laws of thermodynamics they attempt to violate. The continual operation of a machine that creates its own energy and thus violates the first law would be called a perpetual motion machine of the first kind. There is nothing in the first law to preclude turning the heat of the ocean completely into work and hence drive a cargo ship across the ocean. The second law insists that some of the heat utilised be given up to a heat reservoir at a lower temperature. Thus the temperature difference between the top and bottom of the ocean could be used to do work providing some heat is given up to the colder sea water. If not then the second law is violated, and this machine would be an example of perpetual motion of the second kind. The machines illustrated here are nearly all stopped by friction and not by the laws of thermodynamics. Some may apparently work in defiance of the laws but on closer inspection an unexpected source of energy can be found. Device A uses capillary action to overcome gravity but would exchange heat with the atmosphere in the process. The validity of the laws of thermodynamics has been demonstrated by countless indirect scientific experiments, and no perpetual motion machines have been produced that contravene them. A device could conceivably exist that might be kept in motion without violating these two laws, with only dissipative forces such as friction to slow it down. If the device could be made frictionless then its continual motion would be termed perpetual motion of the third kind. Machines B to H all attempt to gain something for nothing using an apparently greater leverage on one side of a device to turn it. The reasoning behind these attempts to break the first law is faulty, as they would remain stationary even without friction. Machines C, D, E, and F had the added dissipative forces of water viscosity or magnetic eddy currents to overcome. Historically thermodynamics was devised for machines containing millions of molecules. Yet single protein molecular machines obey the same laws.

Towards absolute zero

Every substance contains a certain amount of heat, even a relatively cold substance such as ice. The heat is the result of the continual motion of the substance's molecules which, by that motion, possess kinetic energy. Temperature is a measurement of the average kinetic energy of the molecules. The cooler a substance becomes the less its molecules move. Thus it should be possible to continue cooling to the point at which molecular movement ceases completely. This point, "absolute zero", is of great interest to scientists but in practice is unattainable. At temperatures close to absolute zero some materials exhibit remarkable properties, such as superconductivity [6] and superfluidity [Key].

Calculation of absolute zero

On the centigrade temperature scale, absolute zero is 273.16 degrees below the freezing-point of water. Its value can be predicted as a result of the behaviour of gases when they are heated or cooled. When heated, a "perfect" gas expands in volume (V) proportionally to its absolute temperature (T) if its pressure (P) is kept constant. Its pressure increases in the same proportion if its volume is kept constant. The reverse occurs on cooling, according to the equation $PV = RT$ where R is known as the universal gas constant. The pressure actually falls by a factor of 1/273.16 for every 1°C temperature decrease. Thus at −273.16°C zero pressure would be reached and this must be the absolute zero of temperature.

Absolute zero is usually denoted as 0° on the Kelvin scale of temperature, named after the British scientist William Thomson, Baron Kelvin of Largs (1824–1907). Its temperature increments equal those on the centigrade or Celsius scale [1]. Thus 0°K is the same as −273°C (absolute zero temperature is usually rounded to −273°C, or −460° on the Fahrenheit scale) and 273°K equals 0°C, the freezing-point of pure water.

Aiming at absolute zero

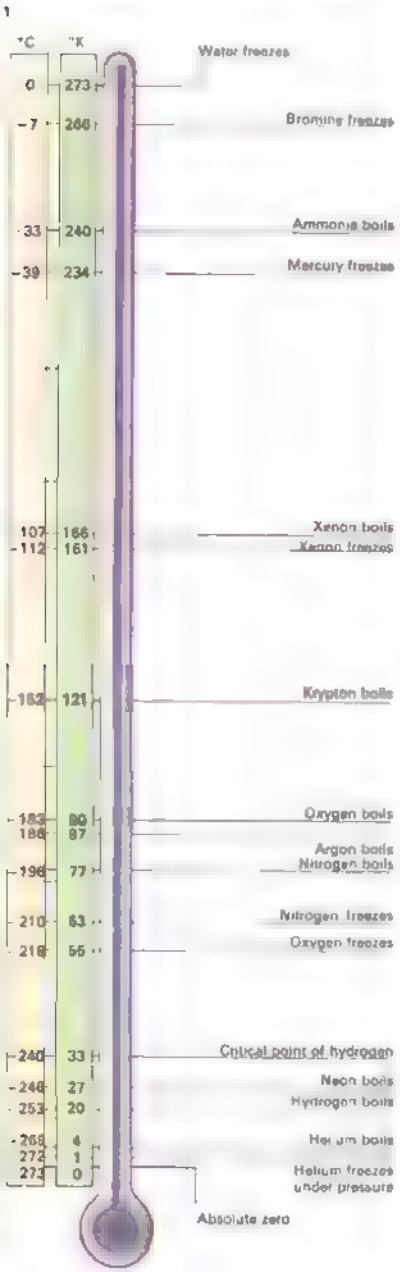
Gas temperatures can be lowered by first compressing the gas in a fixed-volume enclosure and then removing the resultant heat with, for example, a surrounding water jacket. If the gas is then allowed to escape into a larger volume it becomes even cooler because its molecules lose kinetic energy during the expansion. This cycle is used in a refrigerator and can liquefy and even freeze many gases.

The gas most useful in experiments at very low temperatures has been helium, the gas with the lowest boiling-point, 4.2°K (−269°C). The temperature of liquid helium can be further reduced to 1°K by vacuum pumping the gas above the level of the liquid to reduce its pressure and thereby force down the boiling-point. Liquid helium is generally produced in an air liquefaction plant as one of the rare gases left after oxygen and nitrogen have been liquefied [2].

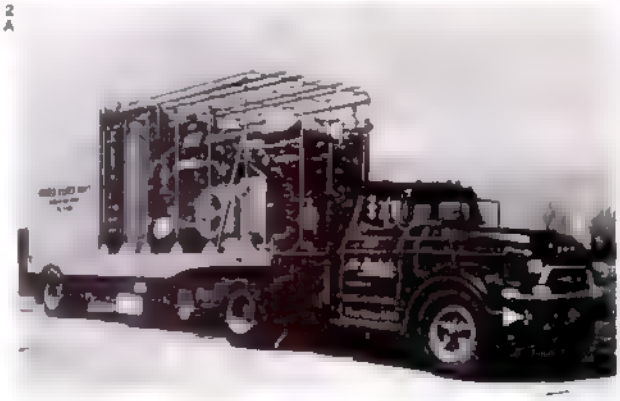
Below 1°K it is difficult to achieve further cooling and a low-temperature effect that occurs in some solids is used [3]. Some salts act as magnets when immersed in a strong magnetic field but stop being magnetic when the field is removed, a phenomenon known as paramagnetism. When the salt is magnetized its molecules line up in the field but are disarranged when the

CONNECTIONS

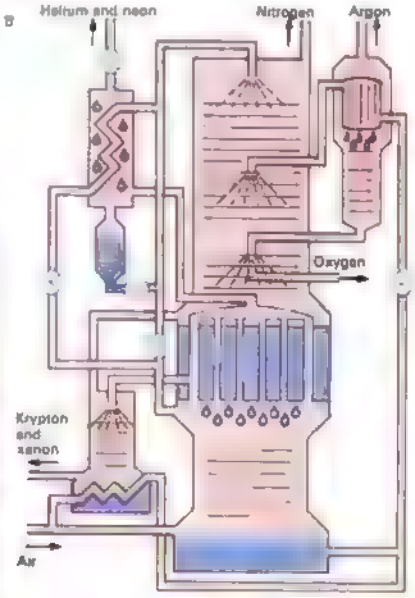
See also

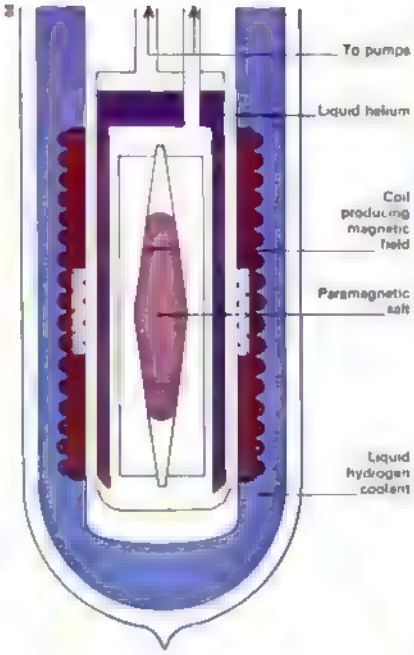


1 The temperatures on this diagram are stated in both the Kelvin (absolute) and Celsius (centigrade) scales, below a temperature equivalent to the melting-point of ice in equilibrium with water.

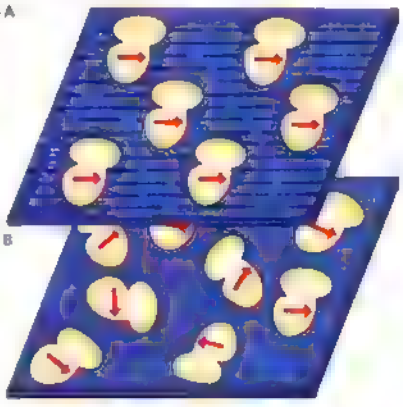


2 In a helium liquefier, which can be carried by road [A], a mixture of helium gas and air is first compressed and the heat generated removed. Air contains, in addition to oxygen and nitrogen, other "inert" gases such as argon, neon, krypton and xenon. At about 20°K all the gases of air except helium can be liquefied in the separator [B]. The helium can be expanded through a nozzle and liquefied





3 To approach absolute zero – below 1°K (−272°C) – a phenomenon called adiabatic demagnetization of paramagnetic salts is used. A paramagnetic salt is one that acts as a magnet only in a strong magnetic field. The field forces the molecules into an orderly array. If there is no exchange of heat energy during this process then it is said to be adiabatic. But when they are ordered the molecules have less energy and the balance is given up to the salt, raising its temperature above 1°K. Liquid helium is then used to re-cool the salt. If it is now demagnetized its molecules become disordered. The energy for this can come only from the salt and its temperature falls below 1°K.



4 The molecules of a paramagnetic salt are normally in continual disordered motion, even if the temperature is as low as 1°K (−272°C). As long as the molecules behave in this way there is no part of the salt that appears to be like either pole of a magnet. As soon as the salt is placed in a magnetic field [A], however, the molecular magnets line up along the field and endow the bulk substance with north and south magnetic poles. Molecular disorder returns when the field is removed [B].

field is removed [4]. If a paramagnetic solid is cooled to 1°K by liquid helium that is allowed to evaporate, heat energy is removed from the solid. When a strong magnetic field is switched on, the molecules align themselves and create heat by their motion. This is removed by the surrounding helium gas, which is pumped away. When the field is switched off, the molecules become disordered and cause a further lowering of the solid's internal energy. The cold salt can then absorb heat from a second helium container. A cycle of magnetization and demagnetization can produce temperatures of a few thousandths of a degree Kelvin.

Superfluidity and superconductivity

Liquid helium at very low temperatures is not only difficult to produce but behaves in a most unusual way. Fast boiling occurs as the vapour pressure falls but at 2.18°K the internal bubbling of helium gas suddenly ceases, although boiling continues. Below this so-called "lambda point" liquid helium exhibits "superfluidic" properties [Key].

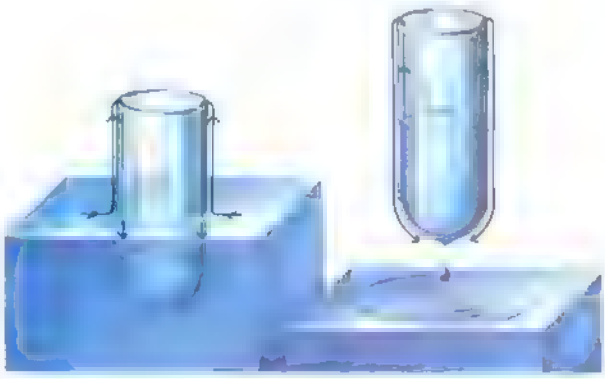
Near absolute zero, certain substances show remarkable properties, for example, a kind of perpetual motion in electric current becomes possible – that is, some metals and alloys exhibit superconductivity [6]. As their temperature is lowered (for example, to 7.2°K for lead) the electrical resistance of the material disappears completely. If an electric current can be made to flow in a ring of such metal it continues to flow. Current has been kept flowing unattenuated in this way for up to several years.

Superconductivity was discovered by the Dutch physicist Heike Kamerlingh Onnes (1853–1926). It can theoretically be used as the basis for some computer memories, for once stored in a superconductor, information remains unaltered. A magnetic field of sufficient strength can destroy the superconducting state and this effect can be used to achieve a high-speed current-switching facility. As a superconducting material has zero electrical resistance, very high currents can pass through it. As a result, superconducting windings for electromagnets can be used to generate extremely powerful magnetic fields.

KEY

Liquid helium cooled below its boiling-point behaves strangely. If a tube is dipped into liquid helium (A) at a temperature of 2.18°K (−270.97°C) an invisible film of liquid creeps up the outside of the tube and then down the inside. The helium fills the tube until the liquid levels are the same inside and out. If the tube is raised a little (B) the "superfluid" helium flows the other way to equalize the levels. Drops of the liquid drip off the bottom of the tube. The thickness of the liquid film can be measured (by light polarization) as about 3 millionths of a centimetre at a height of 1cm above the liquid.

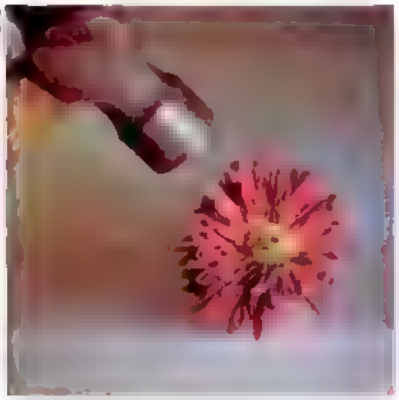
A B



5 Liquid air has a temperature of only 83°K (−190°C) and thus a flower dipped in it will solidify completely as all the fluid in its cells freezes (A). When this happens the flower becomes so brittle that it can be broken into small pieces with a blow from a hammer (B). Liquid air is used industrially for freezing other substances and for the commercial production of oxygen and nitrogen.

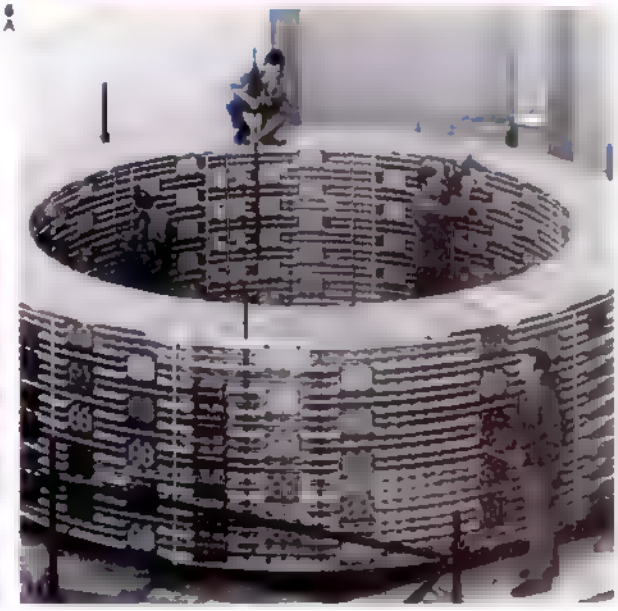
6 The superconducting magnets (A) of a particle accelerator, such as those used in the giant proton synchrotron at CERN in Geneva, Switzerland (B), are products of low temperature physics. Normally superconductivity is destroyed by a high magnetic field. But materials such as niobium-zirconium alloy, with distorted crystal structures, remain superconductive in fields of up to 100 thousand gauss.

5 A

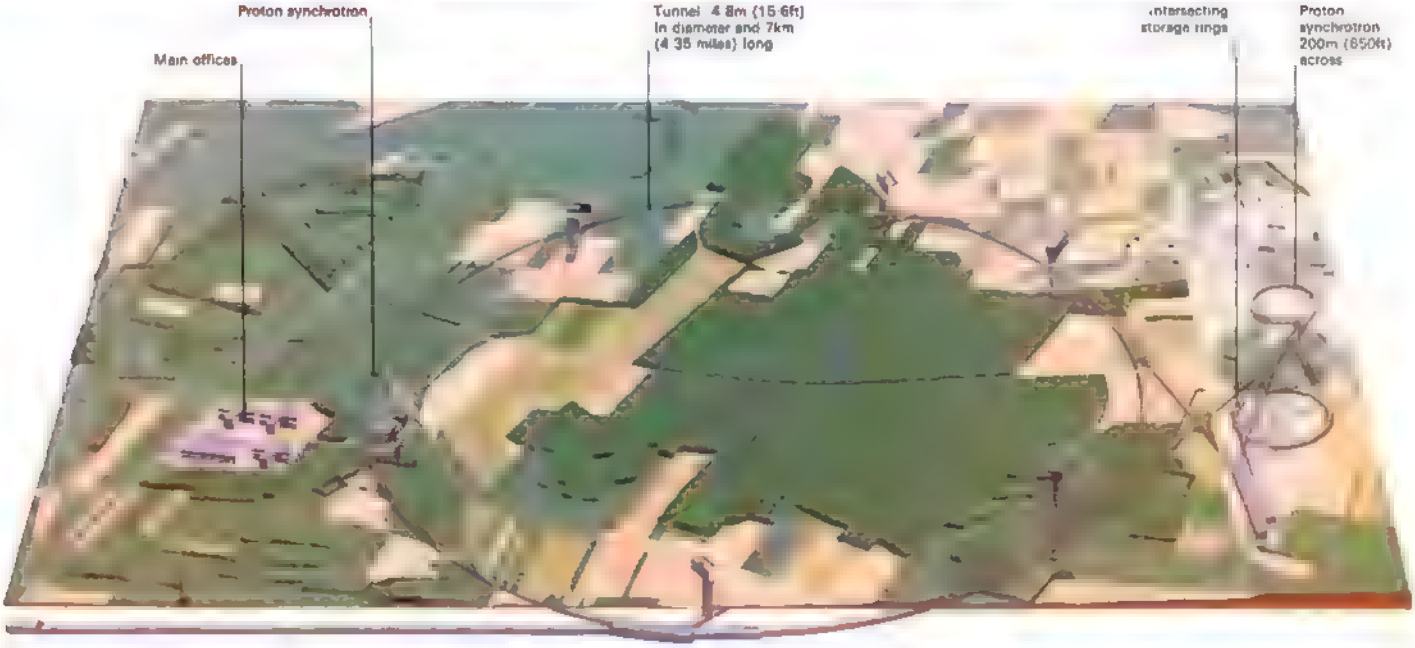


B

6 A



6B



Extremes of pressure

The extremes of pressure – ultra-high vacuum (low pressure) and very high pressure – have varied and sometimes remarkable effects on different materials.

Matter exists as either solid, liquid or gas and all are compressible to various degrees. Perfect gases are compressible to almost any extent, following at low pressures Boyle's law (which states that the volume of a gas varies inversely with its pressure). But liquids are much less compressible and the changes in their volume brought about by pressure follow no simple law.

Solids are the least compressible. Their rigid structure, in which atoms have their mean separation distance fixed by very strong forces, is the most resistant to externally applied pressure. Their structure can be distorted or destroyed by sufficiently high pressures, but the way in which they actually behave is governed by their internal atomic or molecular arrangement.

The compressibility of a gas can be calculated by its equation of state, but that of a liquid or a solid has to be determined experimentally. For the liquid metal mercury at 0°C (32°F), for example, it has been shown that the volume changes by less than one millionth part over a pressure range of 0–7,340 kg/cm² (0–7,000 atmospheres).

Effects of pressure

Pressure applied to any substance can, under certain conditions, cause a change of state. Thus below a certain critical temperature pressure can turn a gas to liquid.

Extremely high pressures find many industrial applications. Hydraulically generated and transmitted pressures are employed for lifting extremely heavy loads. In the motor industry they are used to press completely shaped car body panels [3] from flat metal sheet. The behaviour of metals under compression is also the basis for processes involving rolling and forging. Again, if sufficiently high pressures are brought to bear metal enters what is called its "plastic" range. That means that the metal continues to yield (ie, to extend its dimensions), even though the load applied remains constant. In the normal elastic range, the dimensions change in direct proportion to the applied load and return to normal when the load is removed. The plastic property is used in processes such as extrusion [2], but enormous pressures must be applied to achieve it.

At the other end of the pressure range is the vaccum, which can be of varying degree. To obtain a vacuum, gas atoms or molecules are removed from an enclosed vessel. The number of intermolecular collisions is correspondingly reduced and thus there is a reduction in the internal energy and pressure of the gas.

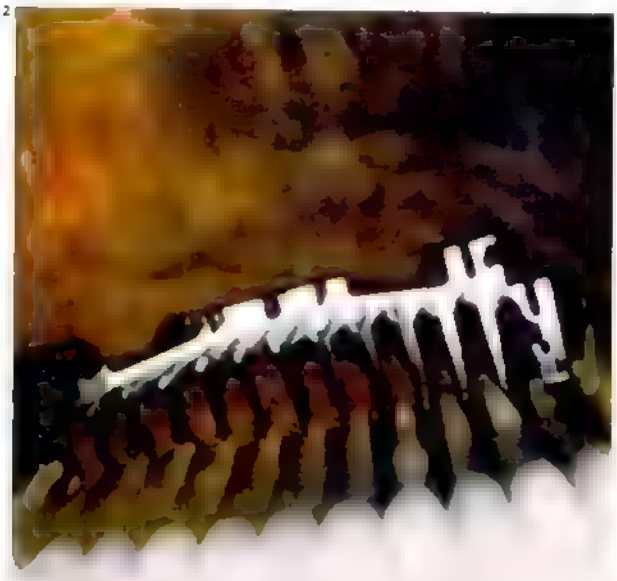
The creation of vacuums is not, however, limited to vessels that previously contained nothing but gas. A partial vaccum can also be achieved above the surface of a liquid – but the space is filled by the vapour of the liquid. As the gas pressure above the liquid is reduced by pumping, the liquid boils at a lower temperature than it does at atmospheric (that is, normal) pressure.

Vacuums in industry

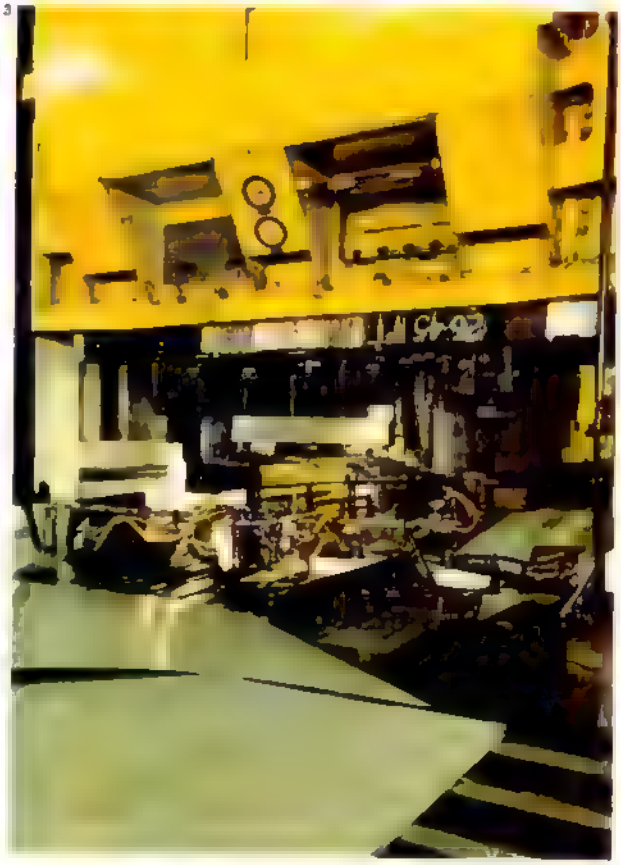
As with high pressures, high vacuum states also have many industrial uses. The process of vacuum deposition by evaporation allows

CONNECTIONS

1 The pressures operating on large structures, such as the bridge spanning the Severn [A], are measured by an electrical strain gauge [B]. This device uses the phenomenom (first noted by Lord Kelvin in 1856) that if a wire is strained its electrical resistance is changed. The principle was first employed practically in the United States in 1938 and now the strain gauge is the most common instrument used for analysis of stress. Gauges can consist of a grid of fine metal wire that is then bonded to a thin backing but, more commonly, a grid of wire filaments is obtained by printing onto a metal foil [C]. In either case the gauge is cemented on to the structural surface and changes in its resistance are measured, the readings are then recorded automatically.



2 Metal forms of complicated shapes can be made by the process of extrusion. Cold metal is forced through a hole of the required shape and size and the great pressures exerted on the metal cause it to assume a "plastic" condition, so that it is able to "flow" smoothly through the extrusion die, as in this machine.



3 Hydraulic presses can make shaped body panels, in one stage, from single flat sheets of metal. A piston carries a former of the shape required and then a hydraulic ram presses this against the metal sheet with tremendous force. The high pressures generally needed call for costly presses, but their speed justifies the expense.

scale objects to be thinly coated with metal [4]. Here, the object to be coated is placed in an enclosed vessel that can maintain a high vacuum. When the coating metal is vaporized within that vessel, it forms a thin mirror-like film over the object.

There are other manufacturing processes in which the controlled deposition of impurities on substances or objects is performed by means of high vacuums. Sophisticated electronic circuitry is based upon this use of an ultra-high vacuum [5].

But technological processes are not the only applications of extremes of pressure. Given that all substances are compressible to some degree, consideration of the effects of increasing pressure on materials to be used in construction work is obviously important. Compression and tension tests of various bearing materials provide the necessary information and increase our knowledge of the extremes of stresses and strains sustainable by different substances.

The behaviour of materials under the two extremes of pressure is well demonstrated in the varying conditions of space. Extremely high vacuum conditions exist in interstellar space, with probably only a few atoms per cubic centimetre. Separate chemical radicals have been found in interplanetary space.

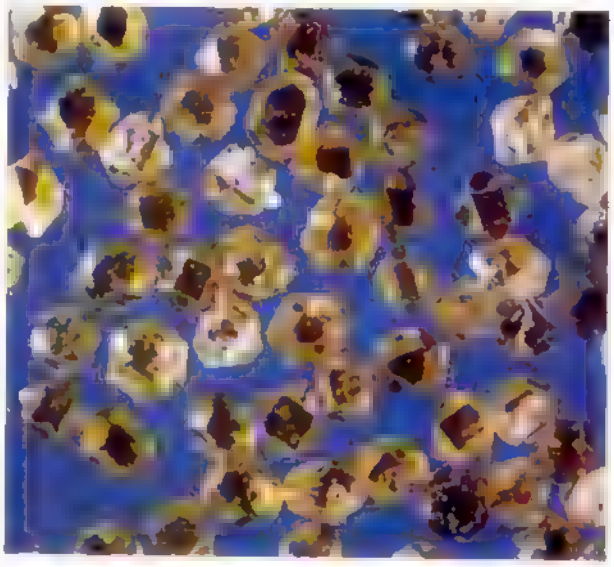
Pressure and gravitation

During the early evolution of stars, clouds of gas and dust (nebulae) condense under the influence of gravitation. Heat is generated, leading to a build up of high internal pressures that tend to resist further gravitational contraction. Compression ceases when these opposing forces of pressure and gravitation are in balance.

Within the stars, the radiation temperatures and pressures are extremely high. At the centre of the Sun the temperature is probably about 10 million °C (18 million °F) and the density fifty times that of water, the pressure amounts to about 400,000 million kg/cm² (6 million million lb per sq in).

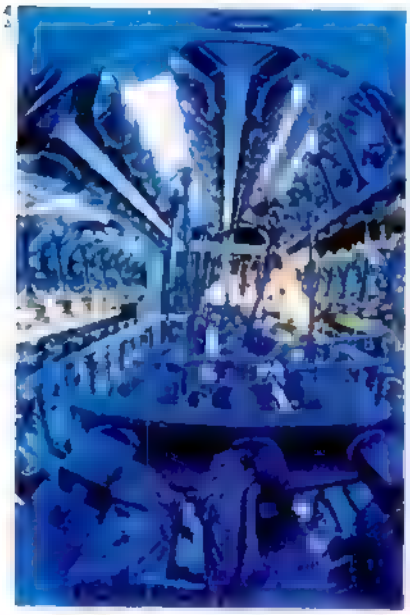
In other stars, the pressures can be so great that the normal arrangement of an atom is completely broken. Protons and electrons are squashed together, creating immense densities of several tonnes per cm

KEY



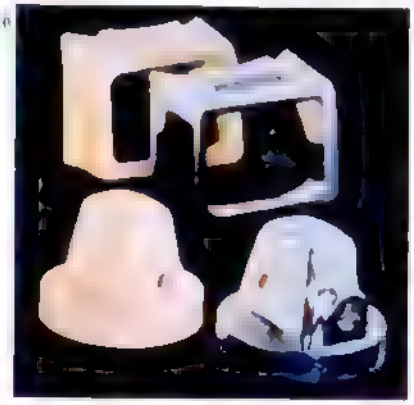
Artificial diamonds are produced when graphite is subjected to very high pressures and temperatures. Most of the stones are inferior to those that have been produced by nature.

4

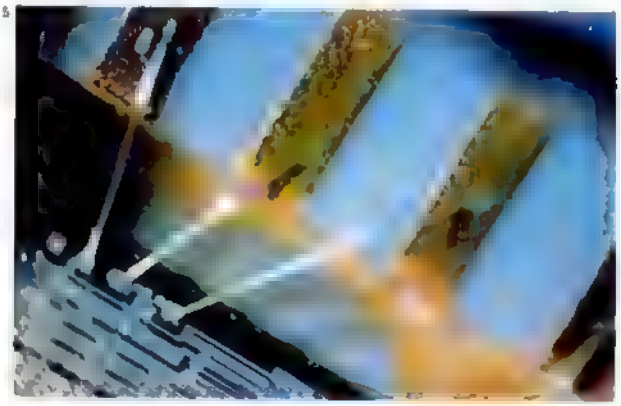


4 Vacuum deposition is widely used to form metal coatings on plastic objects. The object to be so treated is enclosed in a chamber (A), along with a wire filament carrying beads of the metal that is to be deposited. Then the chamber is pumped down to a very low pressure and an electric current is passed through the wire. The beads melt, vaporize and deposit a metal film on the objects. B,

6

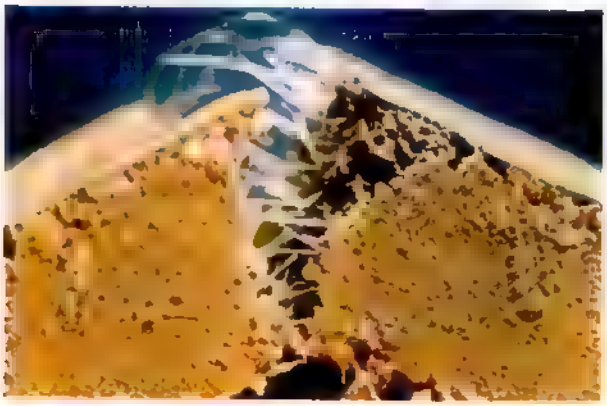


5



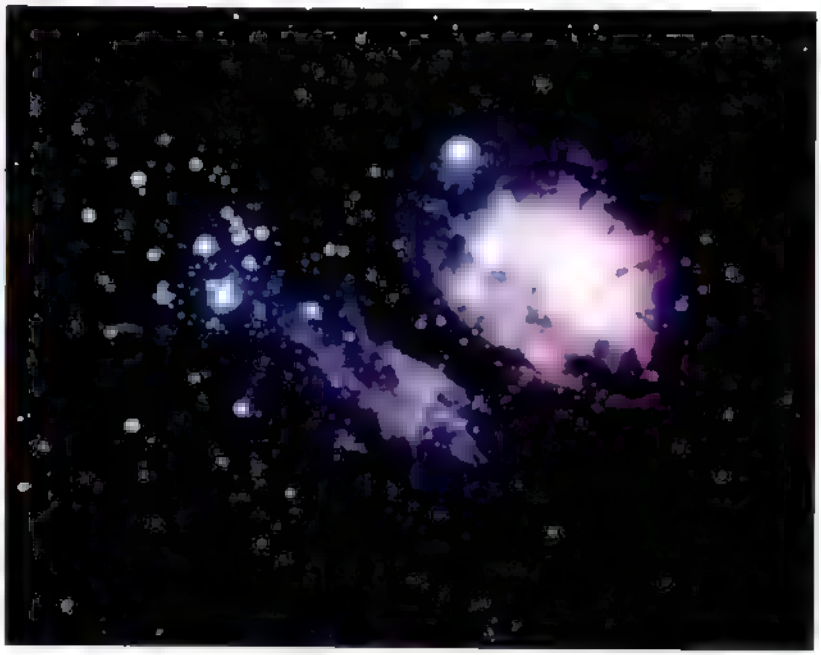
5 Silicon "chips" can be modified by the use of high-vacuum methods so that whole electronic circuits can be integrated into a single piece. To achieve this, certain 'impurities' are diffused into or layered onto the base material. These impurities affect the conductivity of the base and perform equivalent functions to electronic components such as transistors, diodes, capacitors and resistors. The pure silicon base is held in a vacuum during the implantation or layering process. Connections to the circuit are made by fine wires welded to it from gold plated contacts.

8



6 Testing materials to destruction is often necessary to discover their real strengths and their weaknesses. These tests are carried out in a machine that is capable of exerting both compressive and tensile pulling forces and it records automatically the applied pressure on the material at the very instant of its failure. Such tests are not confined to ductile metals (those that can be drawn out as wire) but also applied to brittle materials such as this fibreglass.

7



7 Young stars and a large nebula are seen in the region of Messier 8. In the cluster on the left, star formation appears to be almost complete and very high pressures, accompanied by intense internal heat, prevent the stars from collapsing further.

Light and colour

We are surrounded by various forms of energy, namely light, heat, chemical and mechanical energy. Of these, light is as necessary as heat energy and chemical energy and all are essential to life. Few sightless people can survive without the aid of people who can see. By virtue of its basic nature light enables people to sense the world around them in great detail. This is because light consists of a wave motion of extremely high frequency. If human beings were sensitive instead to radio waves, which have a much lower frequency than light, they would detect no more detail in their surroundings than the blurred outlines seen on a radar screen.

How light travels

Light travels in waves, as sound does. In light the vibrations of the wave consist of vibrating electric and magnetic fields, whereas in sound they are vibrations of a medium such as air or water. Both kinds of waves may vary in intensity, producing stronger vibrations. In sound an increase in intensity causes an increase in loudness and in light it produces an increase in brightness.

Light waves also possess a range of frequencies – the number of vibrations passing every second. In sound, people hear different frequencies as sounds of different pitch, in light they see different colours. Blue light, for example, has a higher frequency than red light. Light may also be considered in terms of wavelength – the distance between successive vibrations in a wave, blue light has a shorter wavelength than red light. Light frequencies are very high and the wavelengths are very short (about 55 millionths of a centimetre).

Paint manufacturers include white and black among their ranges of colours, but these are not strictly colours at all. Black is simply an absence of light and therefore of colour and white light is made up of a mixture of several basic colours. This can be shown by passing sunlight through a prism. The white light of the sun is split up into a band of colours called a spectrum. The spectrum looks exactly like a rainbow [5], which is no coincidence because raindrops act as prisms and split up sunlight to produce a rainbow. The beam of colours from a prism can be made to recombine and produce white light once more. This is a proof that white light is a mixture of all the colours.

The science of spectroscopy

Splitting up light to form a spectrum is important in science [Key]. Different elements glow with different colours when they are heated sufficiently or subjected to an electric discharge – examples are gases in sodium street lights or neon advertising signs. By passing the light from a glowing substance through a prism and examining the resulting spectrum, which has a different pattern for different elements, the glowing substance can be identified. This is useful in all kinds of analysis, but particularly in finding out which elements are present in the Sun and the stars. This branch of science is called spectroscopy.

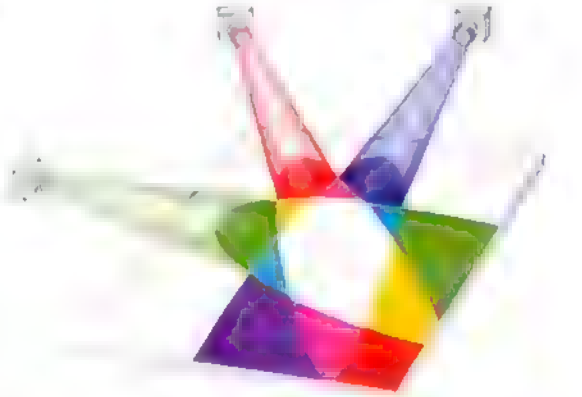
Most of the colour that reaches our eyes comes from objects that are naturally coloured, painted or dyed. When white light strikes the surface of a red object, red light is reflected from it, but all the other colours in the white light are absorbed by the surface. Colour is also produced in other ways. A sub

CONNECTIONS

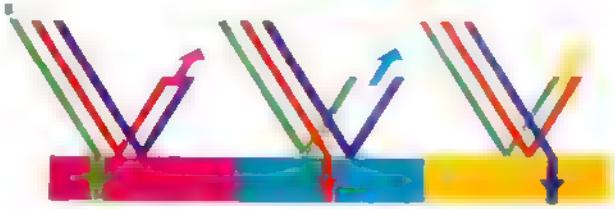
1

1 Colour is used as a means of imparting simple information quickly and unambiguously. On the road, for example, red tail lights on cars indicate "stop" or "danger".

2 A

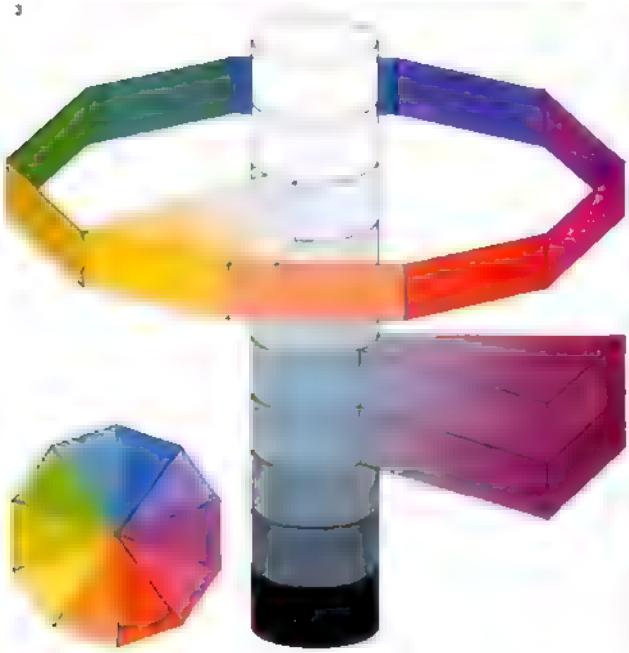


B



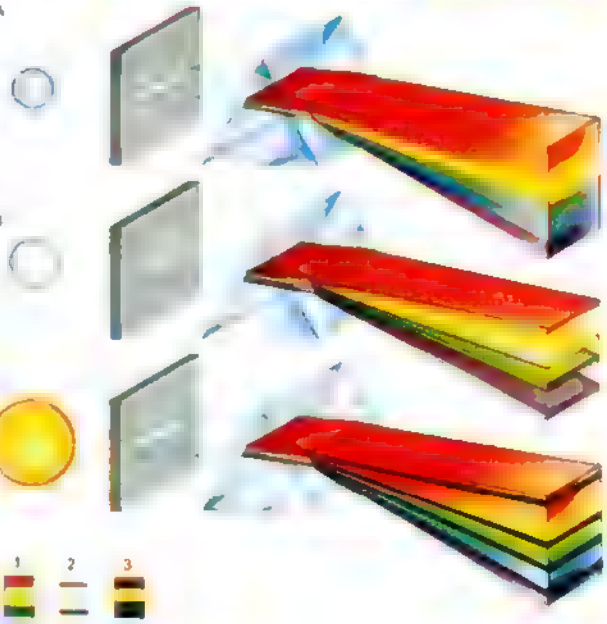
2 Mixing colours depends on whether coloured lights or pigments are being used. Lights combine by additive mixing [A] in which three basic or primary colours, red, green and blue, combine to give white. Yellow, cyan and magenta are secondary colours formed by mixing equally two primaries. Pigment colours combine by subtractive mixing [B] in which some colour is absorbed before mixing of the remaining colours occurs.

3



3 The Munsell colour tree is a system of grading any colour. The hue (basic colour), chroma (amount of colour) and value (degree of lightness or darkness) are measured and the colour's position found among all those in the tree. Hue is denoted by its place on the circumference of the tree, chroma by its distance from the trunk, value by its place up the trunk.

4 A



4 A solid or plasma under pressure heated to incandescence emits a continuous spectrum [A]. At low pressure a gas produces an emission spectrum [B]. In the Sun [C] light from the inside [1] is partly absorbed as it passes through the outer regions [2] to form an absorption spectrum [3].

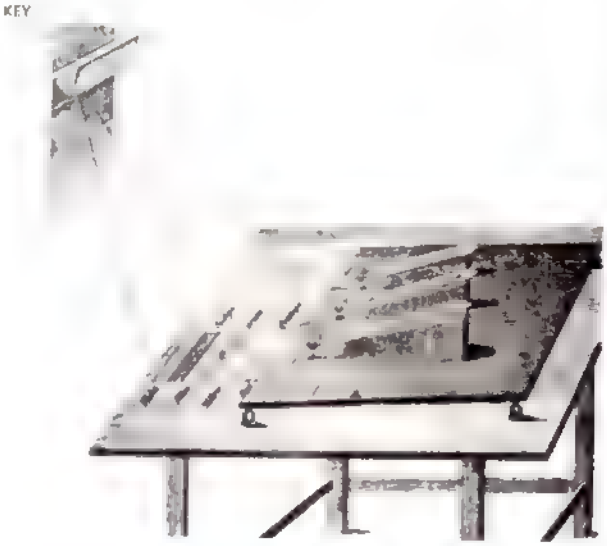
stance can be heated so much that it glows with colour and luminous compounds such as the phosphors in a colour television screen light up with colour when they are struck by invisible cathode rays (beams of electrons) or ultra-violet rays.

Humans (and many other animals) can perceive colour because the retina in the eye contains three kinds of light sensors. These detect different ranges of light frequencies, roughly corresponding to red, green and blue. All other colours can be produced by combining light of these three basic colours in various amounts. Red and green combine to produce yellow; green and blue to give cyan, and blue and red to make magenta. All three basic colours combine to give white light.

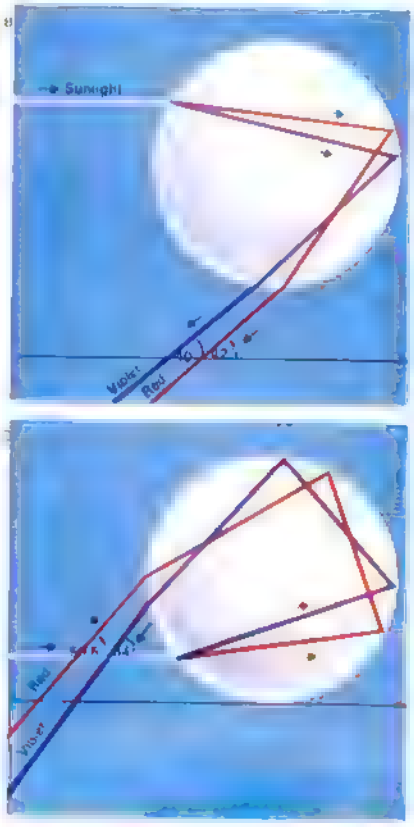
Additive and subtractive mixing

It may seem strange, to anyone used to mixing red and green on a paint brush and obtaining brown, to read that red and green make yellow. This is because coloured lights and coloured paints combine in different ways. A colour television set produces coloured light and close examination of a lit screen reveals that it contains patterns of red, green and blue dots or stripes. At a distance the dots or stripes merge into a colour picture. But close up, the yellow light can be seen to be made up of red light and green light. This kind of colour mixing, in which light combines directly, is called additive mixing [2]. Any three coloured lights, such as red, green and blue, that combine to form any other desired colour or in the right proportions to form white light are known as the primary colours.

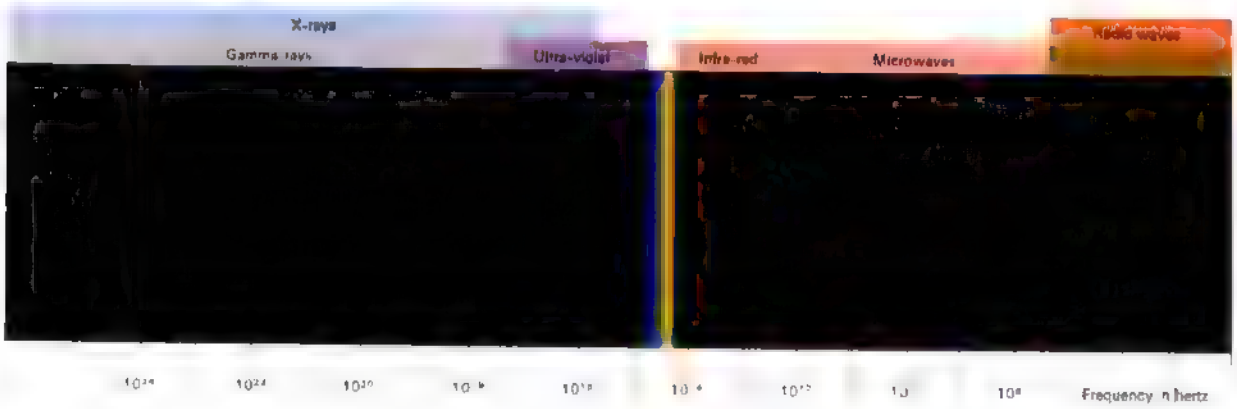
In producing colours by mixing paints, dyes and inks, subtractive mixing occurs [2]. The colours form not by mixing the three basic colours directly, but by absorbing some of them from the light that illuminates the surface of a painted object. Thus yellow paint absorbs blue from the illuminating white light but reflects red and green, which combine to reach the eye as yellow. Cyan paint absorbs red light, leaving blue and green to mix and make cyan. Mixing yellow and cyan subtracts red and blue from the white light, but leaves green to be reflected: so cyan and yellow paints mix to give green.



Dispersion of light by a prism produces a spectrum of colours. In 1800 the British astronomer William Herschel placed thermometers just beyond the red end of the spectrum and observed a rise in temperature. He deduced that the prism was dispersing invisible heat rays (now called infra-red rays).



5 A rainbow is a natural demonstration of the mixture of colours, or wavelengths that make up white light. Drops of moisture in the atmosphere act as prisms, dispersing the light into its component colours. The observer sees the various colours of the spectrum emerging from many droplets. The completeness of the rainbow he sees depends on his position between the Sun and the drops of moisture and on his horizon, from aircraft at high altitudes a circle can be seen. Sometimes two rainbows are visible (A): an inner – called a primary bow – and a larger outer one, or secondary bow. The colours are in opposite sequence in the two rainbows. In the primary (B) light is dispersed on entering a rain-drop and the colours reflected from the back of the drop towards the observer. Some rays may be reflected twice within the drop (C) to produce a reversal of the colours in the secondary.



6 The spectrum of visible light is only a small part of the much greater spectrum of all electromagnetic radiation. Beyond the blue end lie the invisible ultra-violet rays, X rays and gamma rays, while infra red rays (heat rays), microwaves and radio waves lie beyond the red end. All electromagnetic radiation has the power to penetrate matter to a certain extent. High frequency radiation – that is, X rays and gamma-rays – penetrates most

Mirrors and lenses

Nearly everyone looks in a mirror at least once a day and people with less than perfect vision spend most of their lives looking through spectacle or contact lenses. Telescopes, binoculars, microscopes, cameras and projectors help us to examine the world about us in far more detail than can be perceived by the unaided human eye. All these visual aids and optical instruments use mirrors or lenses. They work using the simple laws of optics, but before these laws and their application can be understood it is necessary to appreciate how an image is formed.

Light rays and images

Any illuminated or luminous object sends out light rays that spread in all directions in straight lines. An image forms if any of the light rays coming from the same point on the object happen to meet. Normally images do not form because there is nothing to bend the light rays to make them meet, but a lens will do this. The image produced can be seen on a card or a screen placed at the point where the rays meet. If the rays meet exactly a sharp image is formed and the image is said to be in focus. It is known as a real image and is the kind produced in a camera or by a projector. If the incoming light rays are parallel, the image is produced at a distance called the focal length of the lens.

Plane (flat) mirrors, on the other hand, produce images that cannot be shown on a screen. In this case the light rays are bent but they continue to become farther apart (diverge) rather than closer together (converge). But the human brain always assumes that light rays reach the eyes in a straight line, and we therefore see an image at the point that the object would be occupying if the rays were not bent [6]. This kind of image is called a virtual image and it is always sharp.

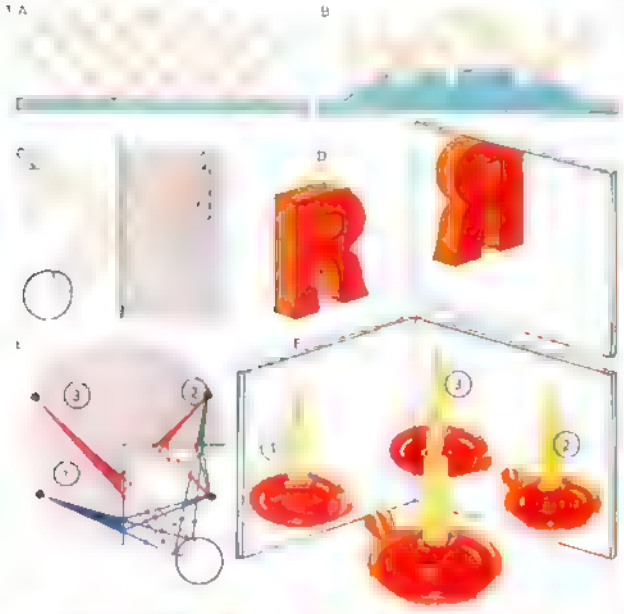
Except when illuminating the deepest black objects, light is reflected from every surface it strikes. A dull or matt surface scatters the rays at all angles. But a very smooth surface acts as a mirror and reflects all rays so that the angle of incidence equals the angle of reflection. A plane mirror, being flat, bends all the rays that strike it by the same angle and so a virtual image, unchanged in size and shape, is seen in it. Left and right are interchanged because a light ray leaving one side of the object is reflected by the mirror to the opposite side of the eye from the one it would otherwise strike [1].

Curved mirrors [2] produce images changed in size and shape. A convex mirror, one that bends outwards towards the observer, gives a smaller virtual image. It makes the rays diverge more than if they came from a plane mirror, producing a smaller virtual image behind the mirror. A concave (inward-bending) mirror may make the rays from the object converge, giving a real image in front of the mirror. If the object is very close, the rays diverge less than with a plane mirror, giving a magnified virtual image behind the mirror.

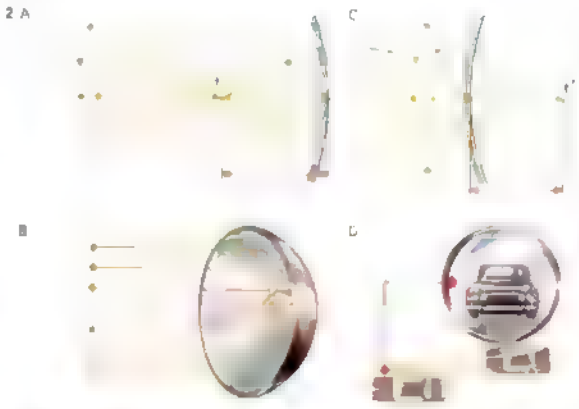
Bending of light

If light rays meet a transparent object most of the rays enter the object and emerge from the other side. The rays are bent as they pass through the surfaces, being deflected away from the surface if they are entering a denser medium and towards the surface if they are leaving a denser medium. This effect is called

CONNECTIONS

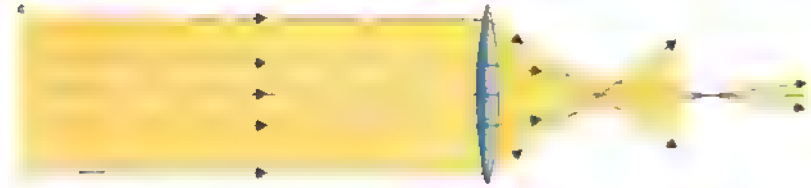


1 A plane mirror reflects all light rays at the same angle [A] whereas a matt surface [B] scatters light. The brain imagines that light rays reaching the eye from a mirror come to it in straight lines and it therefore sees an image at that point where the rays would originate if their paths were not bent by the mirror [C]. The image is seen laterally inverted [D] because the reflected rays reach opposite sides of the eye. An image can be seen right way round in two plane mirrors at right angles [E]. Although an image reflected once appears laterally inverted [1, 2], one reflected twice [3] is seen correctly [F].

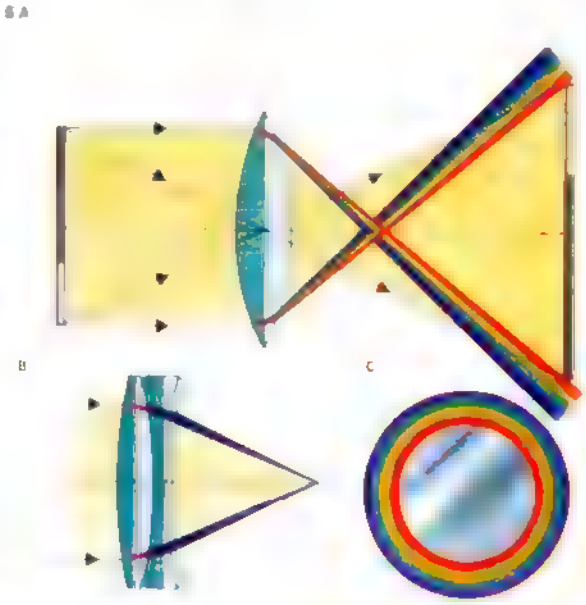


2 Curved mirrors form real and virtual images. A concave mirror [A] reflects the rays of a parallel beam of light so that they converge and meet at the focus of the mirror [f]. A real image of a distant object will be formed on a screen placed at the focus. Also, if a light source is placed at the focus, a parallel beam of light is given. The reflector of a car headlamp works in this way [B]. A convex mirror [C] makes the rays of a parallel beam diverge as if they were coming from the focus [f'], which produces a diminished virtual image, as in a driving mirror [D].

3 Lenses form images because light rays are bent by refraction as they pass through a lens. The rays are made to converge on passing through a convex lens [1] or diverge through a concave lens [2], regardless of the direction in which the rays are moving. Thus convex lenses can make light rays meet and produce real images, whereas concave lenses give only virtual images.



4 Spherical aberration produces a blurred image. Rays passing through the centre of the lens are brought to a focus at a different point from rays passing through the edge of the lens and there is no place at which all the rays come to a single focus and give a sharp image. Spherical aberration may be reduced by narrowing the lens so that rays do not pass through the edges and by combining lenses to cancel out the defects in each kind of lens.



5 Chromatic aberration produces coloured fringes around the lens edges, and parts of the image may not be sharp [C]. This aberration occurs with single lenses because they behave like prisms and bend blue light more than red light [A]. Combining the lens with a weaker concave lens [B] made of a different glass cancels out this dispersion effect, and both red and blue rays are brought to the same focus to produce a sharper, more distinct image.

refraction. Some of the light is reflected instead of being refracted. If, on leaving the denser medium, the light rays strike a surface at or below a certain angle, called the critical angle, they are all reflected back (none is refracted) and the light stays inside the denser medium. The amount of bending depends on the refractive index of the medium; the greater the refractive index the greater the amount of bending

Refraction and lenses

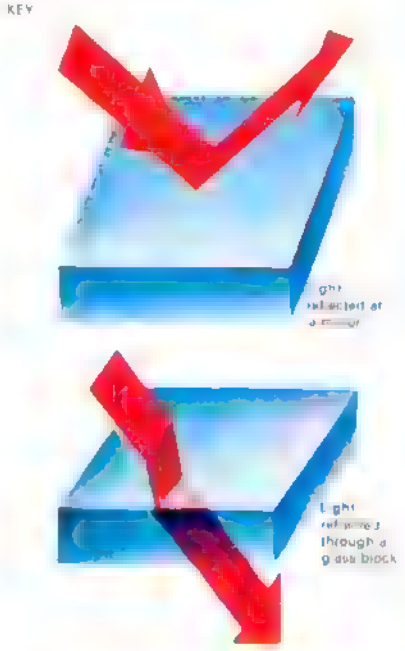
Refraction explains why objects seen in water appear to be less deep than they really are. Light rays from the submerged object bend as they leave the water but the eye, as always, imagines that the rays have come in straight lines from the object

Lenses work by refraction [3] and are shaped to bend the light rays passing through them by different amounts. In a convex lens – one in which the surfaces curve outwards – light rays passing through converge and meet to form a real image The lenses in a camera, a projector and in the eye work in this way But if the object is very close to the lens, the light rays diverge after passing through the lens When viewed from the other side, a large virtual image of the object will be seen behind the lens. This is how a magnifying glass works Concave lenses – which curve inwards – make light rays diverge and produce small virtual images. These lenses are used in spectacles to correct short sight, converging lenses being necessary for long sight

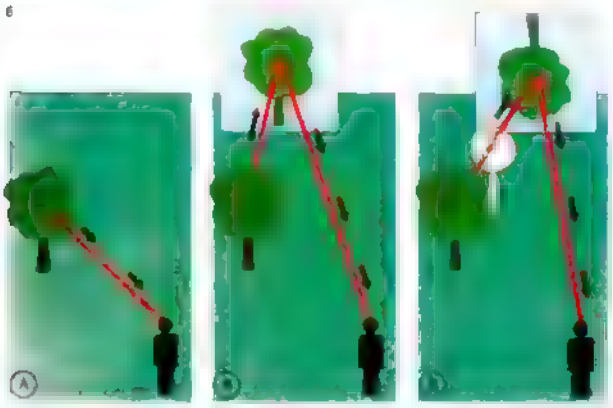
Single lenses produce several kinds of aberrations [4, 5], or distortions, of the image. Coloured fringes may be seen around the edges, and parts of the image may not be sharp When several lenses are combined the aberration is reduced by cancelling out defects in each kind of lens, and high-quality lenses consist of several different elements grouped together to give a perfect image Each element has to be carefully shaped and positioned with great accuracy.

Most optical instruments use lenses to produce images [8, 9, 10]. But the largest telescopes used in observatories all have a concave mirror to produce a real image of a distant planet, and this image is then viewed with a magnifying lens to enlarge it

Mirrors and lenses alter the paths of light rays in accordance with simple laws. A mirror reflects light rays so they leave the mirror at the same angle as they strike the mirror The brain assumes that the light rays reaching the eye have travelled from an object in straight lines, and it therefore sees an image behind the mirror A transparent material refracts light passing through it so that the path of the rays is bent at the surface by a certain angle, which depends on the refractive index of the material The brain assumes that light moves only in straight lines and the image it sees of an object is displaced



6



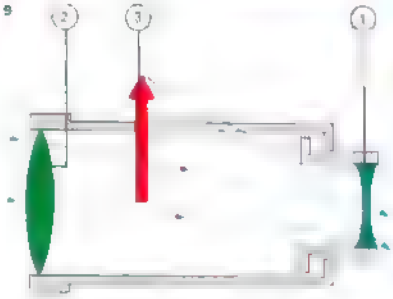
5 Images form when the eye receives light rays coming from the same point on an object and bends them to meet at the retina The tree reflects light rays in all directions. The eye may see it directly [A], see a virtual image of the tree "behind" a mirror [B] or view a real image of it formed on a screen by a lens [C]. The images are located at the points from which the light rays appear to originate

7 A periscope, in its most simple form, consists of two mirrors angled at 45° one above the other so that an image is reflected from the top of the instrument down to the observer at its base. A submarine periscope works on the same principle, but has prisms instead of mirrors and a system of lenses to produce a magnified image or a wide field of view. Optical adjustments can be made with the handles

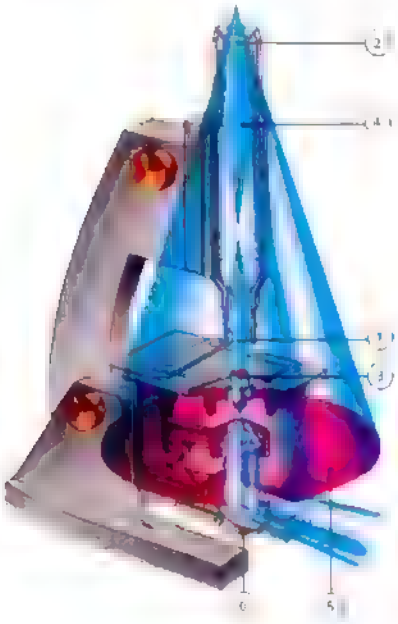
8 In a microscope an image of the specimen is formed by the objective lens and this image is then viewed by the eyepiece lens so that it is magnified and an extremely close view obtained

1 The field lens forming part of the objective
2 The eye lens of the eyepiece
3 Specimen placed on the glass slide
4 The objective of the eyepiece
5 The microscope effectively increases the angle at which light from the specimen enters the eye, and the final virtual image appears to lie in this plane
6 Mirror reflects light on to specimen

9

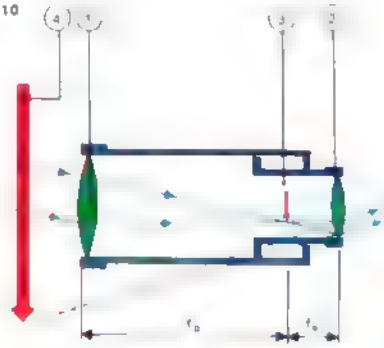


8



9 Opera glasses have a pair of telescopes known as Galilean telescopes. The concave eyepiece lens [1] is placed inside the focus of the convex objective lens [2] to obtain a magnified upright image [3] of the object. The lenses have low-power magnification Galileo discovered the moons of Jupiter using a similar instrument

10



10 An astronomical telescope has a similar optical system to a microscope. It consists of an objective lens [1] (of focal length f_o) and an eyepiece lens [2] (of focal length f_e) Parallel light from a distant object converges to form an inverted image [3] Light rays then appear to come from the large upside-down image [4]

7

Light waves

In the seventeenth century scientists were divided in their opinions about the nature of light. Some believed it to be made up of streams of particles of some kind, while others argued that it consisted of waves. Reflection was easy to explain by the particle theory, it could be pictured as a bouncing of particles off surfaces rather as balls bounce from the sides of a billiard table. (The modern form of this theory envisages light travelling as "packets" of energy called photons.) Refraction was more difficult to explain – why should some particles bounce off and others pass through a surface? Other effects were impossible to account for by the particle theory and the wave nature of light became accepted as more convincing.

The wave motion of light

In those early days no one knew exactly what vibrated in light to make it behave as a wave motion, nor how the waves could be produced. These problems began to be solved in the 1870s. It was discovered that a light wave consists of vibrating electric and magnetic fields travelling through space; the two fields vibrate at right-angles to each other and to the direction of motion. In fact light waves are part of a whole group of electromagnetic waves that include X-rays, ultra-violet rays, infra-red rays and radio waves. Light waves can be produced by changing the orbits of electrons inside atoms. If an atom receives energy in some way – perhaps as heat, light or electrical energy for example – the electrons move away from the nucleus to orbits of higher energy. They then jump to a lower energy orbit and give out energy in the form of electromagnetic waves as they do so. In this way objects produce light [1].

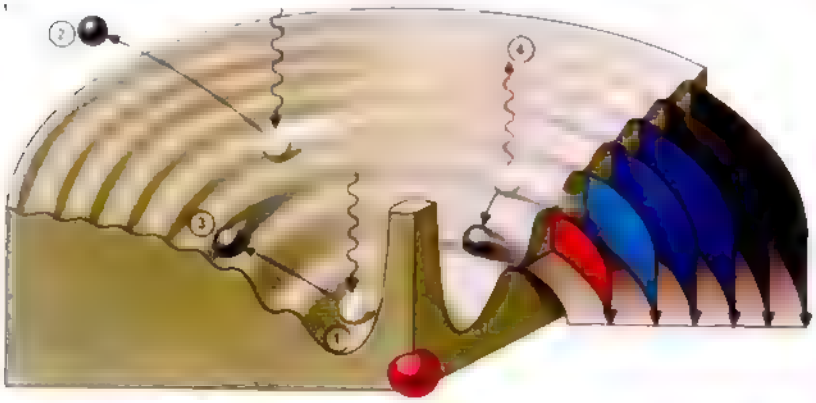
Light spreads out from any point producing it or reflecting it in ever-expanding spheres, rather as ripples spread out in circles over a pond. Each ray of light can be thought of as moving in a straight line, producing a continuous series of ever-expanding vibrational movements through space. In all the rays leaving a point the vibrations add up to give a set of spherical wavefronts consisting of alternate peaks and troughs of energy. Each peak and trough are maxima of vibration but in opposite directions.

The shadow of an object is rarely seen to have sharp edges, but this is because a source of light always has a certain size [Key]. If the source were infinitely small we would expect it to give shadows that are totally sharp because light rays are considered as straight lines, but this is not so. All waves spread round the edge of an object – an effect called diffraction [2]. In the case of light the edge is illuminated and points close to it can act as sources of light waves that spread out in all directions so that the rays are effectively bent by the edge. The wavelength of light is so short that this effect is hard to detect at edges, but it becomes clearly apparent when light passes through very small openings about the same size as the wavelength. This happens in a diffraction grating [3] in which light passes through or is reflected from extremely narrow slits.

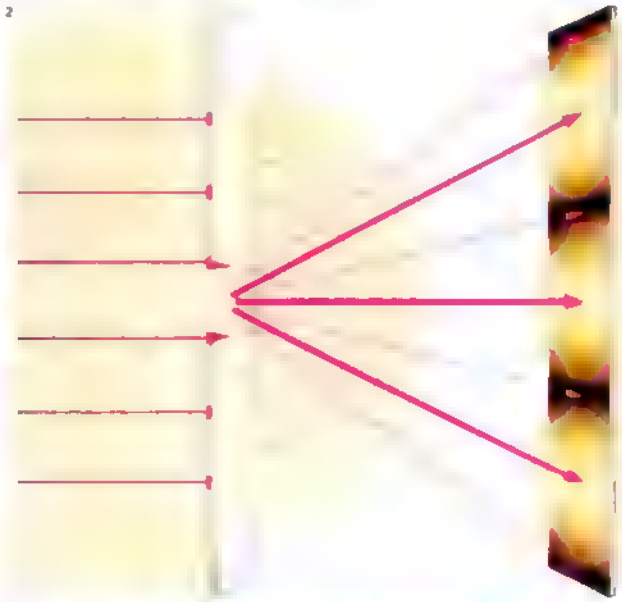
The effects of interference

The wavefronts that spread out from the two edges of the opening cross each other. Where two peaks of the waves meet, an increase in brightness occurs, but where a peak meets a

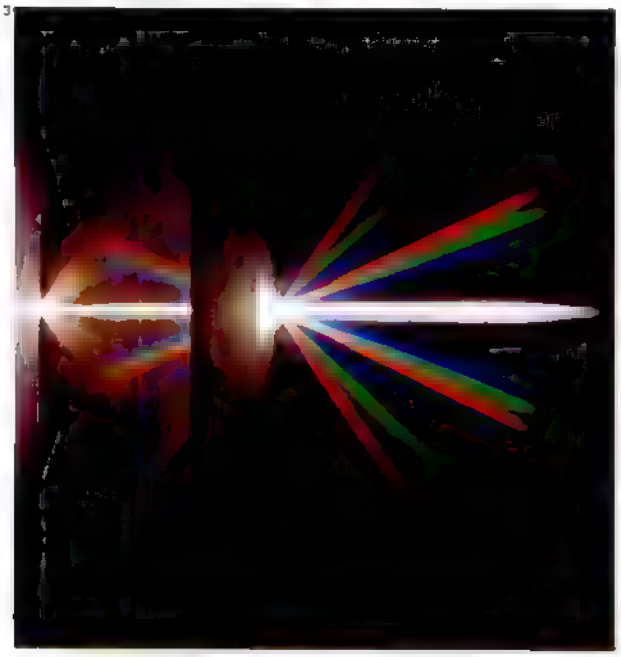
CONNECTIONS



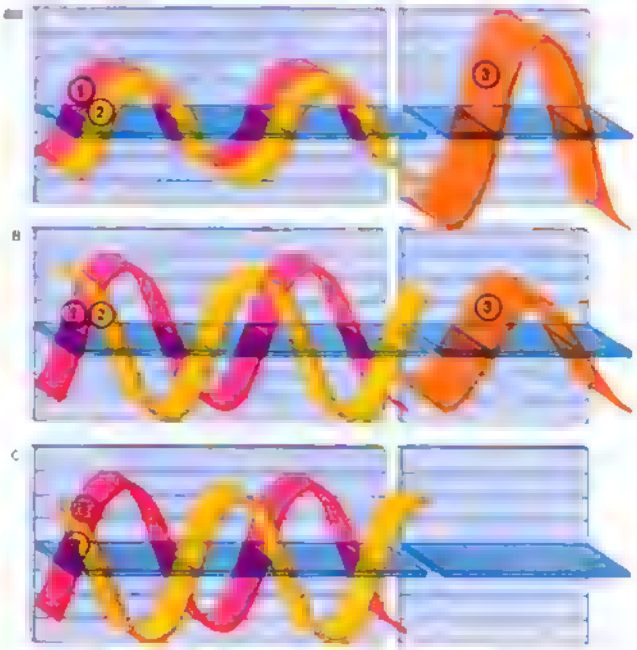
1 Light production can occur when electrons accelerate or when atoms lose energy. In an atom an electron can be pictured as circling the nucleus in a certain orbit (1). If the atom absorbs energy the electron may escape (2) or move to a higher orbit (3). It may then fall to a lower one, giving out visible (4) or invisible radiation depending on the energy gap between the two orbits.



2 Diffraction occurs when a wave passes an edge, causing it to spread around that edge. Often this effect is too slight to be noticed, but it is marked when waves pass through an opening of the same size as the wave length (with light about 55 millionths of a centimetre). The light waves spread out from the edges of the opening and patterns of light and dark are formed where they cross.



3 A diffraction grating has a fine mesh. When white light passes through it is bent in many directions and split up into a spectrum of colours: each wavelength is bent a different amount.



4 Interference occurs when two waves of the same wavelength (1, 2) travel over the same path. The waves interact to give a new light wave (3). If the waves are in phase (A) the new wave is brighter than either of the original waves. If they are slightly out of phase (B) the new wave has about the same brightness as the original waves. If they are totally out of phase (C) the peaks and troughs cancel each other out and so no new wave is produced.

trough they cancel each other out so that no vibration occurs and there is no light. As a result a series of light and dark fringes is produced instead of a single image of the opening [2]. This effect, in which waves reinforce each other or cancel each other out, is called interference [4].

If a ray of light is divided into two rays that later recombine then interference effects are seen if one of the divided rays travels a longer path than the other before the recombination. The peaks and troughs may be out of phase (not exactly 'together') and the light is affected. This happens between two surfaces that are very close together, as in a thin film or two pieces of glass pressed together, and it produces colourful fringed patterns [5]. The iridescent colours seen in the plumage of some birds and some butterflies' wings are produced by the phenomenon of interference, the fine structure of the feather or wing resembles either a diffraction grating or a thin film.

Because interference can be produced by a path difference of only a wavelength or so, interference effects can be used to detect very small changes in length. Interferometers are used for this purpose. They produce interference by dividing a ray of light into two or more beams and then recombining them.

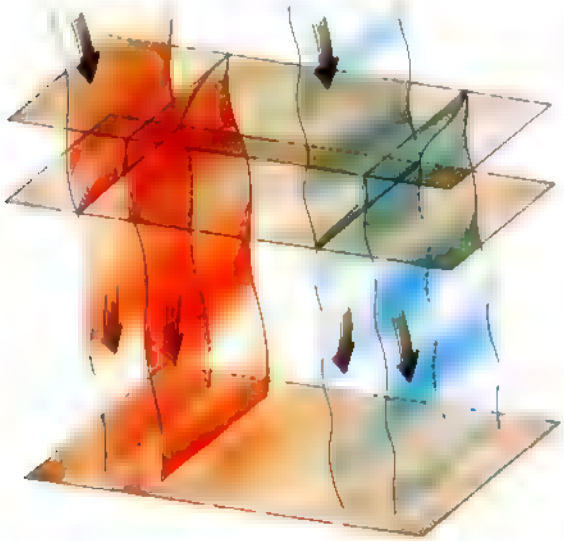
Polarization of light waves

Another effect to be seen with light waves is polarization [7]. In an ordinary light wave the electric and magnetic fields vibrate in many randomly orientated planes about the direction of wave motion; in polarized light they vibrate in only one plane. Light is polarized by passing it through a filter that cuts out all vibrations except those in one particular plane. The polarized beam will then pass through a second filter only if it is set at the correct angle to allow the vibrations through. Otherwise the beam is stopped. Light reflected from surfaces at certain angles is polarized, and polarizing sunglasses [8] cut out glare by stopping reflected beams in this way. Solutions of some chemical substances, such as various sugars, rotate the plane of polarization of light passing through them. The effect is used in chemistry for analysing such solutions.

KEY

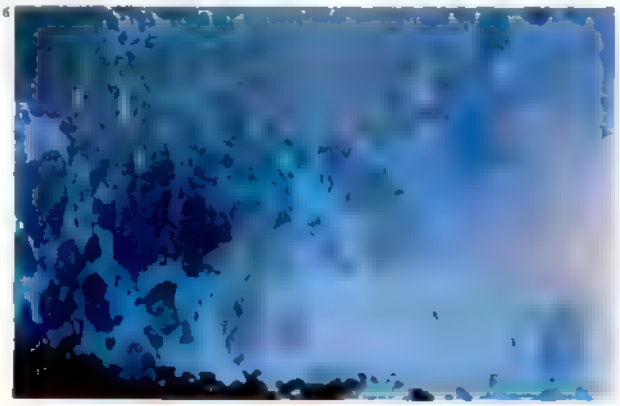
Shadows form with sharp edges when they are cast on a nearby object, but the outlines become less distinct the farther away they are cast. This can easily be explained because light travels in straight rays and every light source has a certain size. The ray paths show that a region exists at the edge of the shadow that is partially illuminated; this region, the penumbra, makes the outlines of the shadow fuzzy. The dark part of the shadow, the umbra, is completely shielded from the light source. The penumbra is less broad the closer the shadow is to the object casting it and so nearby shadows look sharper.

5



5 A thin film, such as a soap bubble or oil film, glistens with colour. Part of the light passing through the film is reflected between the inner surfaces of the film and emerges to interfere with the rest of the light that passed straight through. Travelling paths of different lengths, some of the waves are in phase and reinforce each other (red) while others (blue) cancel each other out and are not seen.

6



6 Interference is responsible for the coloured reflections from bubbles and oil films on water. The light reflected from the top of the film interferes with light reflected from the lower surface.

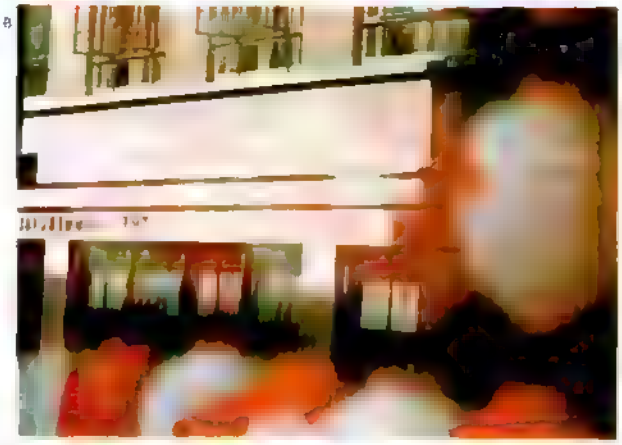
7A

C

D

7 Unpolarized light consists of vibrations in all planes at right angles to the direction of the light wave, the arrows show the wave approaching head on (A). Polarized light consists of vibrations in one plane only (B). Light rays consist of vibrating electric and magnetic fields at right angles (C), only the electric vibration denotes the plane of polarization (D).

8



8 Light reflected from glass or water is partly polarized. Here (B) a reflection makes it difficult to see through a shop window. A similar photograph taken with a polarizing filter over the camera lens gives a reflection-free view (A). Polarizing sunglasses reduce glare this way.

9

9 Stresses and strains in transparent materials such as glass and plastic become visible when viewed with polarized light. Here the regions of strain in a heat treated car windscreen become visible as spectral colours.

The speed of light

Every time we press a light switch, light floods the room instantaneously – or almost instantaneously. It does take a fraction of a second for the light to pass from the light bulb to our eyes but the time taken is far too brief for us to be aware of it. To early scientists light seemed to take no time to propagate, many claimed its velocity to be infinite.

Determining the velocity of light

Others, notably the Italian astronomer Galileo (1564–1642), challenged this view. He attempted to measure the velocity of light by trying to find out how long light took to travel between two hills a known distance apart. His experiment was inconclusive but it did show that if light has a particular velocity then it is very great. Confirmation of this view came with observations of the moons of Jupiter by the Danish astronomer Olaus Roemer (1644–1710) in 1675 [1]. The moons, which had been discovered by Galileo in 1610, are often eclipsed by Jupiter but Roemer found that predictions of the eclipse times were as much as 22 minutes out. Roemer reasoned that the variations occurred because the distance between Earth and Jupiter varies depending on their positions in their orbits around the Sun, and light therefore takes different times to reach the Earth from Jupiter. Knowing the distances involved Roemer made a good estimate of the velocity of light, obtaining a value of 227,000km (141,000 miles) a second. The true velocity is almost 300,000km or 186,000 miles a second.

Another astronomical determination of the velocity was made by the English astronomer James Bradley (1693–1762) in 1728. He observed that the stars are seen in slightly different directions depending on the position of the Earth in its orbit. This phenomenon, called stellar aberration, is caused by the Earth's motion and the differences in direction are simply related to the difference between this motion and the velocity of light. Bradley was therefore able to obtain a value for the velocity of light, and it was of the same order as Roemer's figure.

Later determinations of the velocity were completely terrestrial and sensitive instruments were used to measure very precisely the time light took to travel a known distance. The instruments contain mirrors to reflect a ray of light along a particular path and time its passage by a variety of shutter mechanisms. Modern methods use an electronic shutter capable of very rapid action.

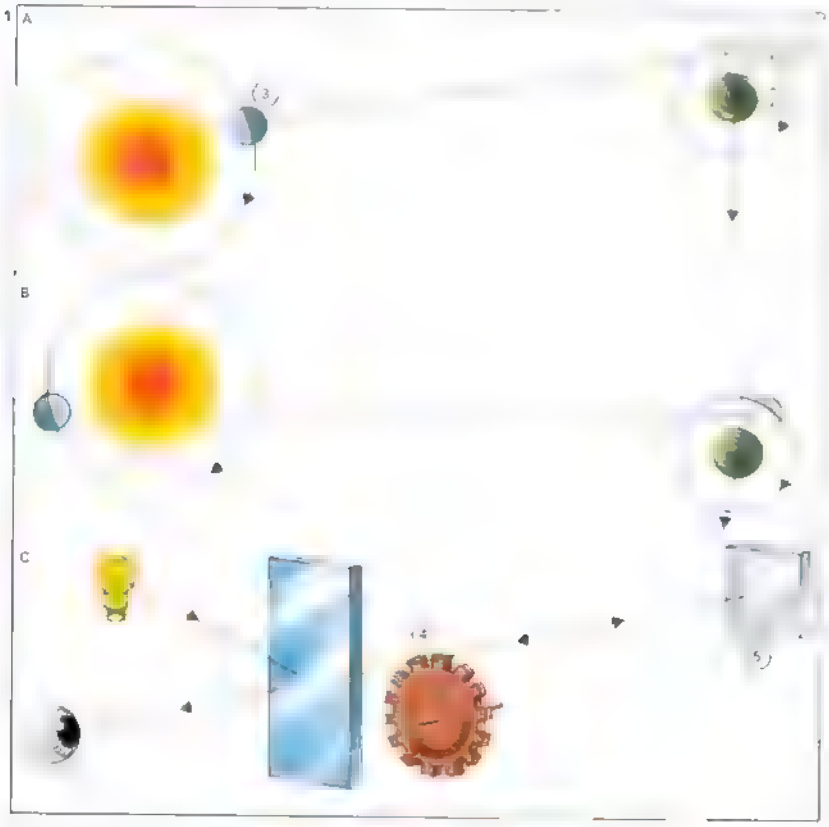
The accepted value for the velocity of light is now 299,792.58km (186,181 miles) a second. This is the velocity in a vacuum, for light slows when it enters a medium such as air, water or glass. The change in velocity causes the light to bend on entering a different medium and refraction occurs. The refractive index of a medium is the ratio of the velocity of light in a vacuum to its velocity in the medium. For example, the refractive index of water is 1.333 or $^4/_3$, and thus the velocity of light in water is only three-quarters of its velocity in a vacuum.

The mystery of the ether

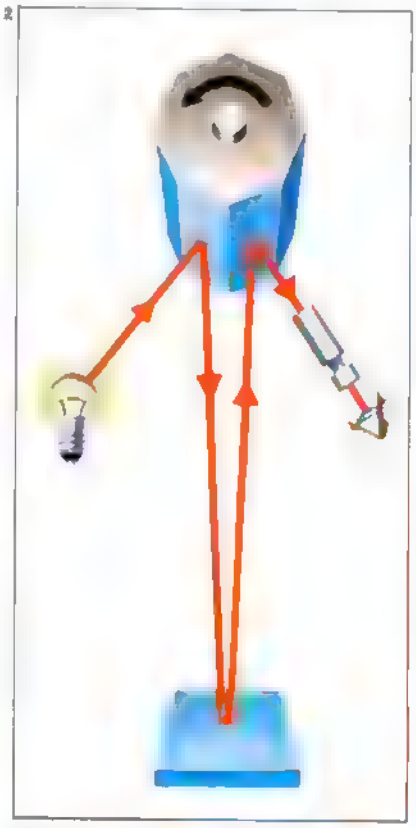
Having determined that light has a certain velocity scientists began to wonder how light waves could travel through space. Other wave motions need a medium in which to travel – sound, for example, moves through air – and light had to have a medium too.

CONNECTIONS

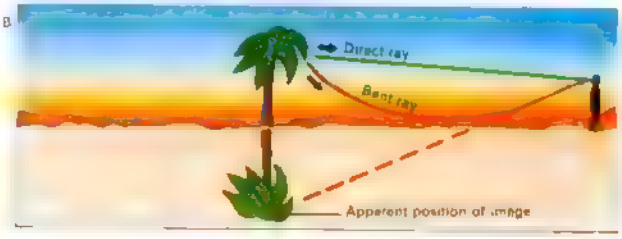
See also



1 The velocity of light was first determined by Olaus Roemer in 1675 [A]. He saw the eclipses of Jupiter's moons [1] by Jupiter [2]. Light from the moons takes less time to reach the Earth [3] when its orbit nears Jupiter than when it is far away towards the other side of the Sun [B]. Knowing the distances and times involved he could calculate the velocity of light. Another determination was made by Armand Fizeau (1819–96) in 1849 [C]. Light was reflected through the teeth of a rotating wheel [4] to a mirror [5] and back through the teeth to the observer. The light was seen only when the wheel spun so fast that no teeth blocked its return journey. From the spacing of the teeth, the speed of rotation of the wheel and the distance of the mirror (8km [5 miles]), the velocity of light could be accurately calculated.



2 A rotating mirror was used by Michelson to measure the velocity of light in 1927. Light travelled from one face of the mirror to a plane mirror 35km (22 miles) away and then back to another face and an eyepiece. An image of the light source was obtained with the mirror first stationary and then rotated at sufficient speed for the image to be seen in the same position. At such a speed the mirror turned so that the next face moved into position as the light made its 70km (44 mile) journey to and from the plane mirror. Velocity of light was calculated from the speed of rotation.



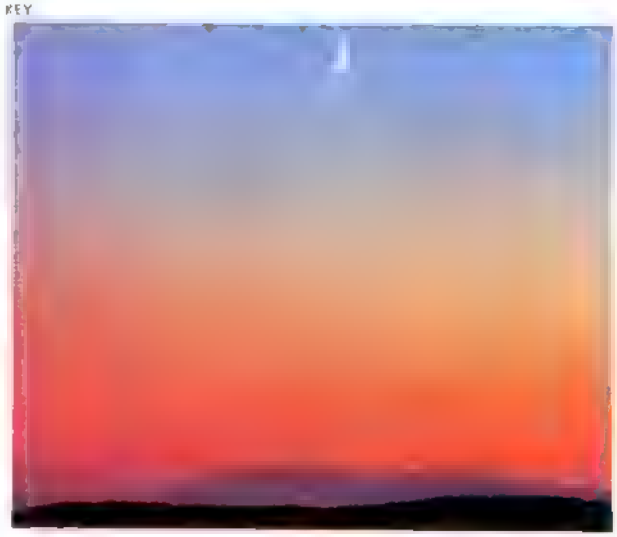
3 A mirage is seen in a desert [A] because the hot air acts like a mirror. The heat of the sand causes layers of air at different temperatures to lie above the ground and rays of light from the ground bend as they move through the layers [B]. Each layer has a different refractive index. In extreme conditions the light rays bend so much that they are deflected back towards the ground and an image of an object over the horizon is seen.

As the medium through which light moved could not be seen to exist, one was invented, it was called the ether and it was supposed to pervade the whole universe. Thorny problems surrounded the ether. Known wave motions move more rapidly in denser, more elastic substances and a wave motion as fast as light should theoretically need a medium denser than steel. Yet the planets continue to sail through space, unimpeded by the ether. There were many other contradictions and so an experiment was made to detect the motion of the Earth through the ether.

In the 1880s two American physicists, Albert Michelson (1852–1931) and Edward Morley (1838–1923) [5] made a simple instrument to detect the ether. In it a beam of light was split into two beams at right-angles and the two beams reflected from mirrors before recombining. Combined beams show interference effects if one travels a slightly longer path than the other. Michelson and Morley observed the combined beams in one direction and then turned the instrument at right-angles and observed the beams again. If the light were travelling in an ether it would have to move over a different path in the direction of the Earth's motion from at right angles to it. Turning the instrument at right angles should show a difference in the interference effects if the ether existed. None was observed and none has been observed in many repeats of this classic experiment.

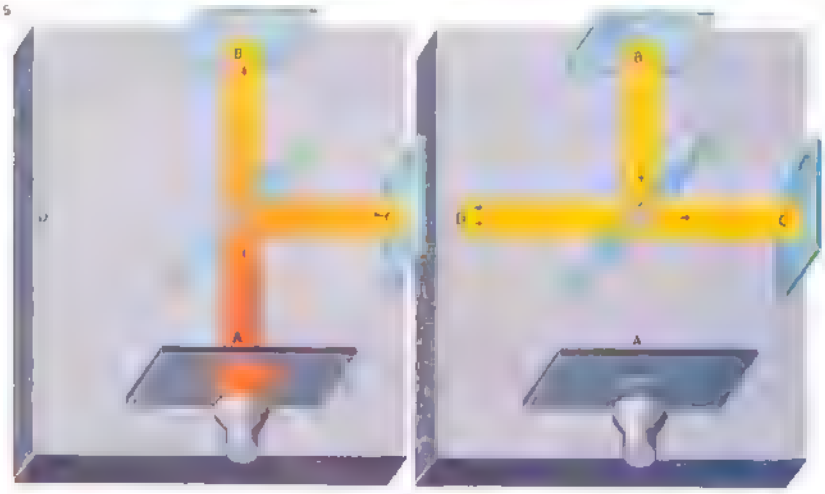
The basis of relativity

The conclusion of the Michelson–Morley experiment was that ether does not exist and light does not need a medium for its propagation, or that the ether can never be detected. Without a stationary ether there is no basis in the universe against which the absolute motion of everything can be measured, except for light. The Michelson–Morley experiment showed that the velocity of light is the same in the direction of the Earth's motion as at right-angles to it and is always the same whatever the observer's motion. These conclusions had profound implications but to realize them it took a genius – Albert Einstein (1879–1955) – who used them as a basis for the theory of relativity.



Light that reaches us from heavenly bodies does not travel instantaneously. It takes 1.25 seconds to get to Earth from the Moon, 8 minutes from the Sun, over an hour from Saturn and the outer planets, and over four years from the nearest star. We see the galaxies as they were millions of years ago.

4 The blue glow coming from the water surrounding this nuclear reactor is known as Cerenkov radiation. It is produced because nuclear particles emitted by the reactor are moving faster than light does itself in water, which slows light by about a quarter. The particles cause a shock wave to be produced in the water just as a supersonic plane produces a shock wave in the air. We hear the sound shock wave as a supersonic boom and see the light shock wave as blue light. The production of Cerenkov radiation is used as a method of detecting fast-moving particles in nuclear physics. The radiation was first observed by the Russian physicist Paul Cerenkov (sometimes known as Cherenkov) in 1934.



5 The Michelson-Morley experiment, first made in 1881, used an interferometer to produce a pattern of interference fringes from two beams at right-angles. The Earth's motion was expected to make the light move faster along one path (ABD) than along the other (ACD) so that on turning the interferometer a change in pattern would be seen. In the experiment no change was detected.

6 The motion of an observer does not affect the velocity of light but it does change the colour or wavelength of light to give the Doppler effect. However, only stars move fast enough to show the effect. If the star and the observer are moving apart (A) the wavelength increases because the individual waves are encountered less frequently. The light is more red than if the star were stationary (B) and this red shift shows as a shift in the lines of the spectrum of the star from (2) to (1). If a star and observer are approaching (C), wavelength decreases and the light appears bluer (3). A similar shift of frequency is used in radar to detect the movement of aircraft or cars.

The idea of relativity

Relativity sought to eliminate from physics the idea of absolute values for space and time. Such values were held to be fixed and quite independent of the person measuring them or of the instruments used. To Isaac Newton (1642–1727) they existed as a backdrop against which he could formulate general "laws" about such quantities as acceleration and force. It was the genius of Albert Einstein (1879–1955) that, through the special and general theories of relativity, showed that such absolutes did not exist and that Newton's laws were not universally true.

The special theory of relativity

Einstein's special theory of relativity (1905) was based on the idea that all uniform motion is relative – that is, an object can be seen to move uniformly only in relation to some stationary frame of reference. The classic experiment made by Albert Michelson (1852–1931) and Edward Morley (1838–1923) determined that the speed of light is always the same in a given vacuum regardless of the speed of the source of light, of anyone observing it or of its wavelength [1, 2]. From these results Einstein deduced an astonishing set of conclusions. They showed that the mass, length and time interval of an object will appear to change when the object begins to move relative to an observer.

If, say, an astronomer were to observe an extremely fast-moving spaceship, then his instruments would indicate that the mass of the spaceship had increased, that all lengths in the direction of the spaceship's motion had decreased, and time aboard was slower. Yet in the spaceship itself nothing would appear to have changed, although if the pilot looked back at the astronomer – who would be in the same motion relative to him – he would observe that mass, length and time there had changed in exactly the same way.

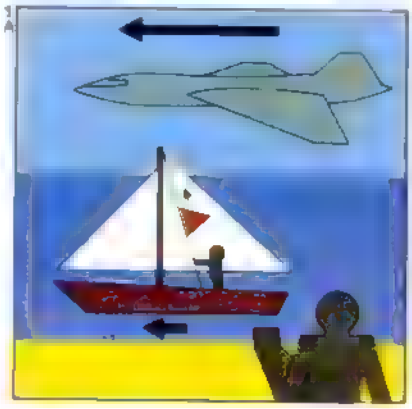
The light clock [4A] shows why time varies with motion and by how much. Normally the effects of special relativity are undetectable in an object until it is travelling at nearly the speed of light (300,000km [186,000 miles] a second) [3], although very sensitive atomic clocks have been used to detect clocks "going slow" on aircraft in flight. The effects do become large for subatomic particles moving at close to the speed of light. Thus, because of their high speed very fast unstable particles in cosmic rays live longer in the Earth's atmosphere than would otherwise be expected [4B]. Sub-atomic particles can be so speeded artificially that their masses are increased many thousand-fold, particle accelerators have to be specially designed to allow for this effect.

It is Einstein's famous equation "$E = mc^2$", relating the energy E and mass m of a moving particle with c the velocity of light that shows why in special relativity a particle given ever greater energy will increase its mass. Because c^2 is so large, only a small amount of mass is equivalent to a vast amount of energy. The conversion of mass into energy takes place in nuclear reactors, in atomic power stations, in nuclear weapons and in the Sun and other stars.

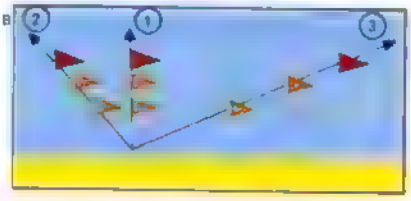
As the speed of a particle approaches that of light its energy increases indefinitely. But there is a limit to the amount of energy available to any particle and so it can never travel faster than light. The light barrier cannot be crossed, but there may exist parti-

CONNECTIONS

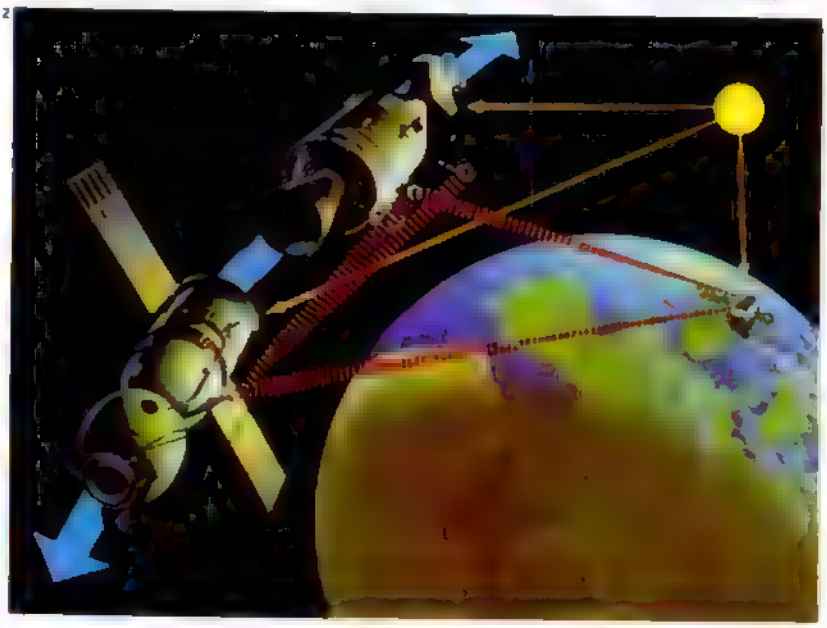
See also



1 Relativity hinges on the simple idea that all motion is relative. A sailor in a yacht hauls a pennant up the mast [A]. To him, it appears to move vertically up [1]. To a man on the shore the pennant appears to move forwards and up [2], because it is being carried past him as it is raised. A passenger in a passing aircraft sees the pennant disappearing rapidly behind him as it is raised [3]. Each observer records the same motion differently [B], none is any more "correct" than the rest, for the planet on which all this happens is also moving. Their views confirm the relativity of all motion.

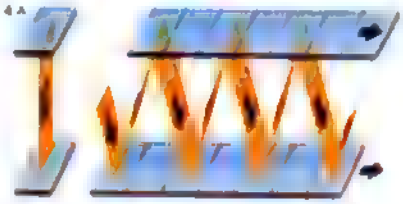


2 The special theory of relativity states that all uniform motion is relative and that the velocity of light is always constant. When two spacecraft pass each other in orbit, each travelling at 8km (5 miles) a second as measured by radar at the tracking station below, the pilots detect that they are travelling relative to each other at 16km (10 miles) a second. If the spacecraft and the tracking station then measure the velocity of light from the Sun, they all get the same result. The spacecraft moving towards the Sun does not get a value that reflects its motion relative to any other body.



3 Einstein's first theory states that the measurement of mass, length and time depends totally on the relative motion of the measuring instrument and the object being measured. Compared with measurements made at rest, the mass will be increased, length decreased in the direction of motion, and time will be slowed. The effects are apparent only at extremely high speeds. At 90% of the velocity of light, mass more than doubles, length reduces by over a half and a clock takes an hour to record 28 min. At the velocity of light, mass would become infinite, length zero and time would slow to a complete stop – a possible situation which means that nothing can overtake the speed of light.

4 A light clock [A], in which light is reflected to and fro between mirrors, shows the principle of time slowed by motion. As the clock and its mirrors move, the light travels farther between reflections than when the clock is stationary, and therefore takes longer to reach the mirrors – observed time is slowed. Time slowing was proved [B] when particles created by cosmic rays (purple tracks) and predicted from observations at high altitudes to have lives shown by the red tracks, reached the ground in unexpectedly great numbers, shown by the yellow tracks.



cles that are always travelling faster than light. These particles, called tachyons, have been looked for but not yet found.

The general theory of relativity

To take account of acceleration [8] and of the force of gravity, Einstein's general theory of relativity (1915) incorporated the fact that all bodies fall equally fast at the surface of the Earth. In other words, the effect of the Earth's gravitational field is an intrinsic feature of the space around the Earth. Einstein described this feature in terms of the curvature of space: the greater the distortion the greater the gravitational force. If time is included with space in this distortion it is possible to incorporate the idea that all motion is relative. The amount of space-time distortion caused by massive bodies can be quantified and it was Einstein's genius that showed how the amount of this curvature depends on nearby massive bodies.

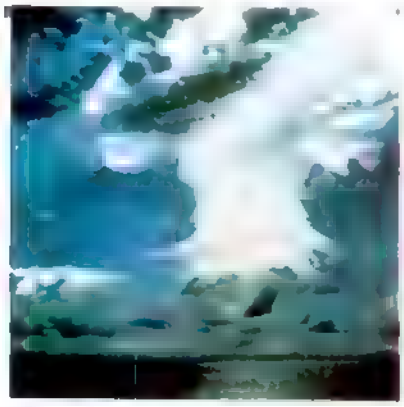
Experimental observations, for example of small deviations in the motions of planets from those predicted by Newton [7], make the general theory of relativity the most satisfactory of a whole range of similar theories. Confirmation also comes from the bending of the path of a ray of light near a massive body. Light has energy – and hence mass – and therefore moves in a curved path in the distorted space around the body [9]. Such bending of light by the Sun was confirmed at an eclipse [10].

Holes in the heavens

All these effects involve weak gravitational fields and cannot put general relativity through the most searching test. When stars have used up their nuclear fuel they may evolve into extremely condensed objects in which strong gravitational fields occur and so they are good testing grounds for general relativity. It is postulated that very heavy stars collapse in on themselves so completely that the escape velocity on their surface is greater than the speed of light. As a result, nothing can ever escape from them again – not even light – and so they are known as "black holes". Good candidates for black holes in our own Galaxy are the variable X-ray stars such as Cygnus X-1.

KEY

Albert Einstein, working without the aid of a laboratory or university post, thought out the revolutionary concepts of the first theory of relativity from simple and seemingly unconnected ideas. Einstein was 26 and working as a patent officer when he published his special theory of relativity in 1905. Ten years later he announced his general theory of relativity. Einstein's fame was world wide, but this did not stop the Nazi rulers of his native Germany from persecuting him for his Jewish blood. From 1933 he lived in the USA.



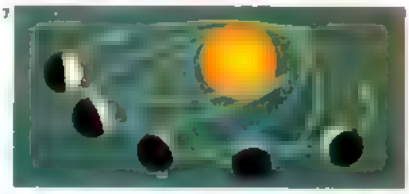
5 Nuclear weapons are one consequence of Einstein's discovery that mass can be converted into energy. But so too are atomic power stations and our understanding of the Sun's energy.

6 Low velocities accumulate by simple arithmetic. If a tank moving at velocity V fires a shell that leaves the gun at velocity v, then the shell will be travelling at $V + v$ [A]. Addition of velocities near that of light (c) is different. If a hypothetical body moving relative to Earth at 0.5c had a supergun that fired a shell at 0.5c, the shell would appear from Earth to move at only 0.8c [B].

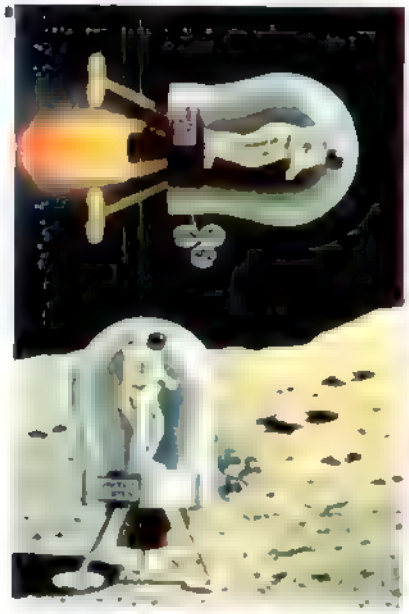
6A

B

7



7 The orbit of Mercury puzzled astronomers because its perihelion (point of nearest approach to the Sun) continually shifted more than could be accounted for by the influence of the other planets. Einstein's general theory of relativity accounted for this movement. He explained that gravity distorts space so the orbits of the planets do not follow the simple orbits described by Newton.



8 The principle of equivalence on which Einstein based his general theory of relativity states that gravity cannot be distinguished from acceleration. An astronaut is pulled to the floor of his stationary craft by gravity (bottom). In the same way as the floor is pushed towards him when the craft accelerates (top). The effects are identical, if he let go of an object it would "fall" in either situation.

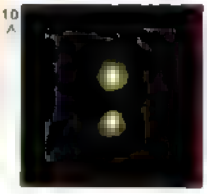
9 A ray of light passing a rotating wheel bearing a line of people would appear to be straight to an outside non rotating observer [A]. As it passes the people are carried away from it by the movement of the wheel. To them the ray appears to bend [B]. This analogy shows that light bends in an accelerating system and therefore, by equivalence, in a gravitational field.

9A

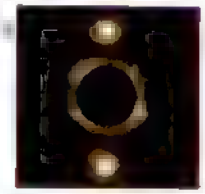
B

10

A

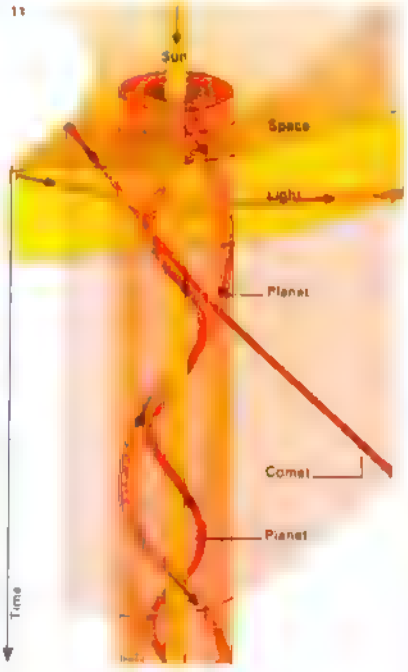


B



10 The bending of light by gravity was detected by photographing two stars normally [A] and in a solar eclipse [B]. As the light rays pass the Sun, they are bent by its field of gravity. As a result the two stars appear to be farther apart [D] than usual [C].

11



11 The dimension of time is as necessary to describe the location of any body as are the three dimensions of space. Einstein realized that if light always travels at the same speed then space and time must therefore be equivalent. This diagram shows the Sun, planets and a comet moving in time as well as space. The varying velocity and widely changing path of the comet demonstrates the effects of the various gravitational fields on its motion, as Einstein correctly predicted in his general theory of relativity.

C

D

Light energy

Light is energy and, in systems of constant mass, energy cannot be created but only changed from one form to another. Light can therefore be produced only from the conversion of some other form of energy. Electrical energy is changed into light in an electric lamp or discharge tube; heat is converted into light in a fire or a red-hot poker; chemical energy is changed into light in luminous animals such as glow-worms. The conversion may also go the other way – light produces electrical energy in a photoelectric cell.

Radiation and quantum theory

The conversion of energy involving light puzzled scientists at the end of the 1800s. A perfectly black object absorbs all light falling on it and all the invisible radiations such as ultra-violet rays and infra-red rays. However, when it is heated, the object gives out radiation, but only at certain definite colours or frequencies. Like the poker in the fire, it first gives out infra-red rays (which can be felt as heat rays), then it glows red, yellow and finally white as it gets hotter. If it could be heated hot enough, it would glow blue-white and emit ultra-violet rays, as do the hottest stars. The wave theory of light [1] could not explain why a black object should differ in the radiation it produces when heated. According to the wave theory, all frequencies should be produced when the object is heated, not different ranges of frequencies at different temperatures.

In 1900, the German physicist Max Planck (1858–1947) put forward a convincing, although revolutionary, theory. He suggested that all energy, including light, consists of whole units of energy: an object can have one unit or a million, but not 0.8, 2.5 or 354.67 units, for example. Each energy unit is called a quantum of energy, from the Latin for "how much". The amount of energy in a quantum is minute and we are unable to make out the individual quanta in light rays as they strike the eye. A quantum of light energy is called a photon.

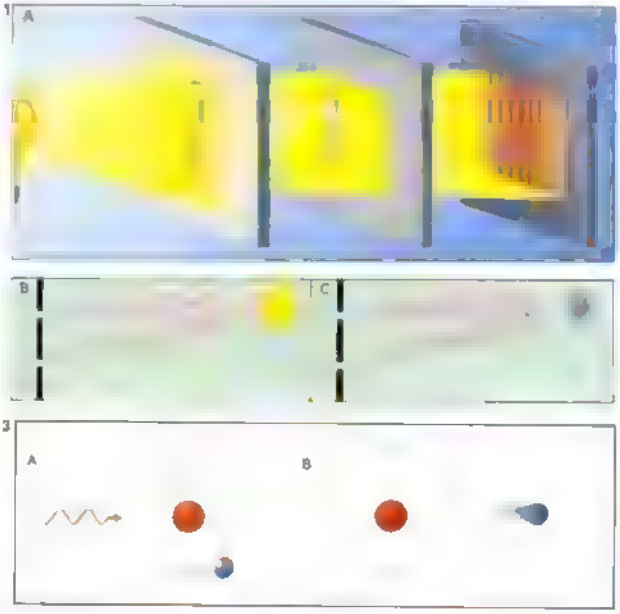
The quantum theory therefore explains why a poker glows in the way it does. As more heat is applied to the poker, the light produced has more energy, and this is shown by a change in colour – a "blue" photon has more energy than a "red" photon. Planck explained that the energy content of each photon of light depends on its frequency, the higher the frequency (more towards the blue or ultra-violet), the greater its energy.

Particles and waves

The idea of light existing in indivisible units was a return to the particle theory of light. A light quantum is included in the fundamental particles that make up matter. If light consists of streams of particles, then it will cross empty spaces with no need for the medium of "ether" that scientists had sought in vain. But such effects as diffraction and interference could be explained only if light behaved as waves. Scientists solved this by assuming that light can behave both as particles and as waves, depending on the situation. This was not just an easy way out of a difficult problem, because the duality can be shown to exist – both experimentally and mathematically. Also, fast-moving particles were found to have wave-like properties. A beam of electrons acts as a wave in an electron microscope, for example.

CONNECTIONS

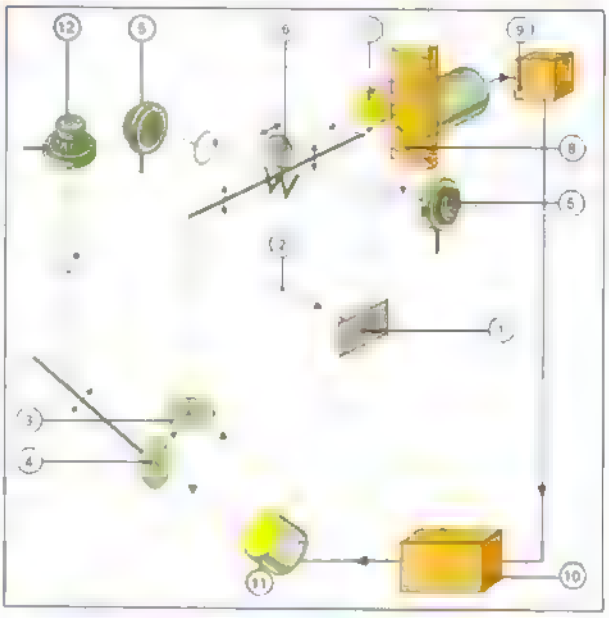
See also

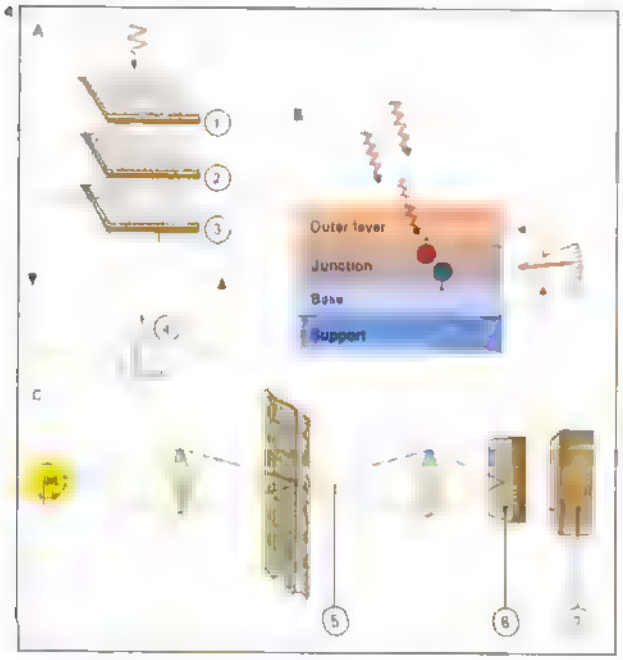


1 The wave nature of light can be demonstrated by passing a narrow beam through two slits [A]. On reaching the screen the resulting two beams overlap to produce a pattern of light and dark stripes – interference fringes. Where two wave peaks, or two troughs, reach the screen in step the combined intensities produce a bright line [B]. A dark line [C] results from a peak and trough arriving together and cancelling out one another.

3 The photoelectric effect was explained by Einstein in 1905 as the absorption of a quantum of energy [A] by an atom and the resulting emission of an electron [B], which can form an electric current.

2 Light exerts pressure on any object it encounters. A vane [1] struck by light [2] moves under this pressure, and the movement is counterbalanced by a horizontal mirror [3] onto which light is directed by a vertical mirror [4]. The torsion heads [5] first level the balance arm. The mirror [6] reflects light from a lamp [7] onto a scale [8]. A timer [9] detects motion of the light across the scale and adjusts a power source [10] to change the intensity of the lamp [11] illuminating the mirror [4] and thus keeps the arm in balance. Deflection of the mirror is detected by the torsion head [12].





4 A light meter [A] contains a glass honeycomb [1] and a grid [2] through which light travels to a light sensitive metal element [3]. There it causes electrons to be excited and these pass round the circuit to the meter [4]. The solar cell [B] is another application of photoelectricity. It consists of layers of semiconductor, usually silicon. Light crosses the outer layer and produces electrons at the junction. These are emitted and travel towards the base, producing a current. The optical sound track [C] on a cine film transmits a varying light signal [5] to a photoelectric cell [6] which produces an electric signal that goes to an amplifier and loudspeaker [7].

5 A photographic exposure meter measures the light coming from a scene. The light strikes a photoelectric cell which produces an electric current that varies in strength according to the intensity of the light. The current is low but sufficient to move a needle across a dial and give a value for the light. Many single-lens reflex cameras have built-in exposure meters that measure the light entering the lens of the camera.

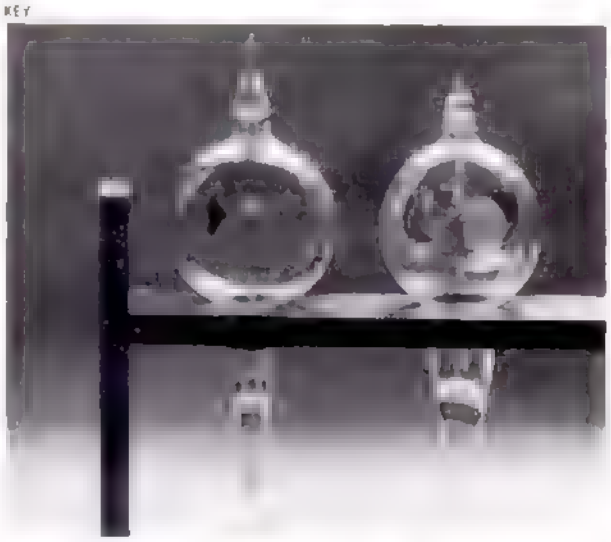
The quantum theory – especially its application to light – finally resolved the problem that had divided scientific thought for centuries. Isaac Newton (1642–1727) had championed the particle theory and Christiaan Huygens (1629–95) had maintained that light travels as waves. With Max Planck's proposal the dilemma ceased to exist – light can be regarded as behaving as particles *or* waves, depending on the phenomenon being investigated.

Certain metals emit electrons when light falls on them – a phenomenon known as the photoelectric effect [3]. It had been observed that brighter light produces more electrons than dim light but not electrons of greater energy; whereas blue light always gives electrons of greater energy than red light, regardless of the intensity of the light. In 1905 Albert Einstein (1879–1955) explained that each electron is released by one photon of light; a bright light has more photons of the same energy than a dim light, but a blue light has photons of greater energy than red light.

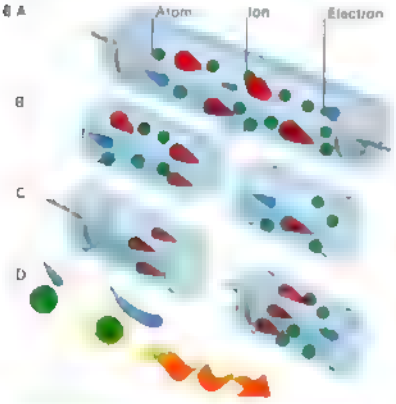
Changing the frequency of light, or converting an invisible frequency into a visible one, has several uses. Fluorescent substances take up light of several frequencies and immediately radiate them at a different frequency, making the resulting colour very bright because extra light has been transformed into it [9]. The fluorescent paints and inks used in some advertisements work in this way. Some washing powders contain optical brighteners that convert invisible ultra-violet rays into blue light and thus making the washing look brighter.

Effects of phosphorescence

Phosphorescence is similar to fluorescence but the production of light continues for some time after the initial radiation has ceased. Television screens contain phosphors that glow for a short time after being struck with the electron beams inside the cathode ray tube and give a picture on the screen. Many instruments make use of the light produced by phosphors to detect invisible rays such as X-rays and fast-moving particles such as cosmic rays. Some phosphorescent paints store light for a long time after being exposed to it and glow in the dark.

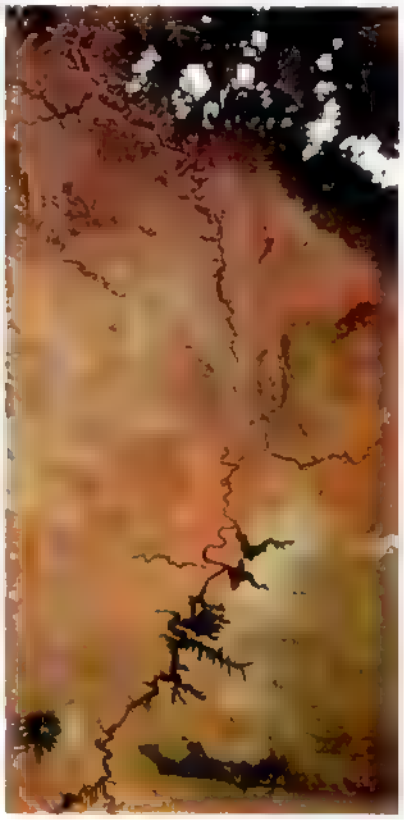


The radiometer, invented by William Crookes (1832–1919) measures radiant energy. When sunlight falls on it, the vanes move round, seemingly pushed by the light. In fact, heat is absorbed by the black side and the few gas molecules left in the bulb's vacuum rebound faster, exerting pressure on that side.

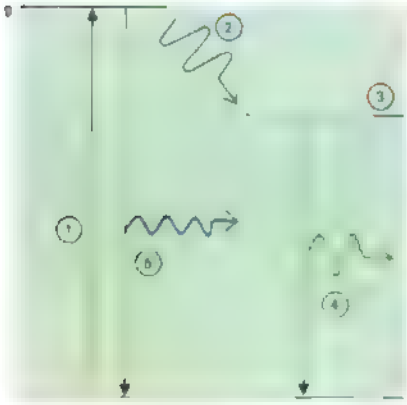


6 A discharge tube contains a gas at low pressure through which electricity is passed [A]. Electrons (negative particles) and ions (positive particles) move towards the electrodes [B]. Ions strike electrodes to produce more electrons [C]. Light is produced as electrons collide with gas atoms [D].

7 Advertising signs contain a gas such as neon or are coated with phosphors to give various colours.

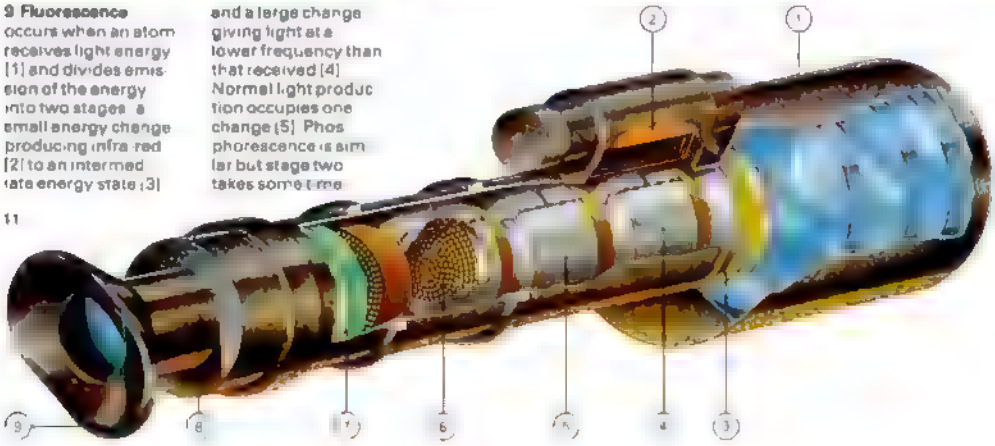


8 An infra-red view of the Colorado River and Lake Powell taken by a satellite shows vegetation in various shades of red and water as black. Diseased plants can be detected by colour.



9 Fluorescence occurs when an atom receives light energy [1] and divides emission of the energy into two stages: a small energy change producing infra-red [2] to an intermediate energy state [3] and a large change giving light at a lower frequency than that received [4]. Normal light production occupies one change [5]. Phosphorescence is similar but stage two takes some time.

10 Automatic letter checking utilizes ultra-violet light. Invisible phosphor codes printed on the stamp glow as the letter passes an ultra-violet scanner classifying it as first or second class.



1 Night lens system
2 6.75 volt mercury battery
3 Photocathode
4 15kV channel plate
5 30 kV channel plate
6 45kV channel plate
7 Fluorescent screen
8 Ocular lens system
9 Eyepiece

11 An image intensifier gives a bright picture of a dimly lit scene. Light from the scene is focused onto the photocathode, which emits electrons. These pass to the electron multipliers that contain tubes lined with electron-emitting substances and produce more electrons from each electron entering the tubes. The resulting intensified electron beam is focused on a fluorescent screen which is viewed through an eyepiece.

Energy from lasers

From boring holes through diamonds to performing delicate eye operations, from spanning space between the Earth and the Moon to detecting the smallest movement, the laser has found an amazing range of uses during its short life. Its future looks no less extraordinary with the promise of three-dimensional television and cheap nuclear power. Clearly the laser is no ordinary source of light.

What is a laser?

A pulse laser is basically a device for storing energy and then releasing it all at once to give a very intense beam of light. The heart of the laser is a crystal or tube of gas or liquid into which energy is pumped [1]. This is usually done by surrounding it with a device to produce a powerful flash of light or an intense beam of radio waves or electrons. As pumping occurs, more and more of the atoms inside take up energy and are excited to high energy states. Suddenly an atom spontaneously returns to its first energy state and gives out a particle of light (a photon). This photon strikes another excited atom and causes it to produce another photon. Very rapidly, a cascade of photons develops. The crystal or tube is closed at both ends by mirrors and the photons bounce to and fro between them building up the cascade. A proportion of this light is able to escape through one of the mirrors, which is half-silvered, and an intense flash of light emerges from the laser.

The first pulse laser, invented by Theodore H. Maiman in 1960, contained a ruby crystal and produced a short flash of red light. Continuous wave lasers now produce continuous beams of many colours and some give out infra-red rays or ultra-violet rays.

The activities of photons

The atoms that discharge photons are stimulated to emit them by the arrival of other photons, which make up light radiation. The light that is pumped into the laser consists of many frequencies but what emerges is a far more intense light at a single frequency.

The result - *light amplification* by the stimulated *emission* of *radiation* - gives the laser its name.

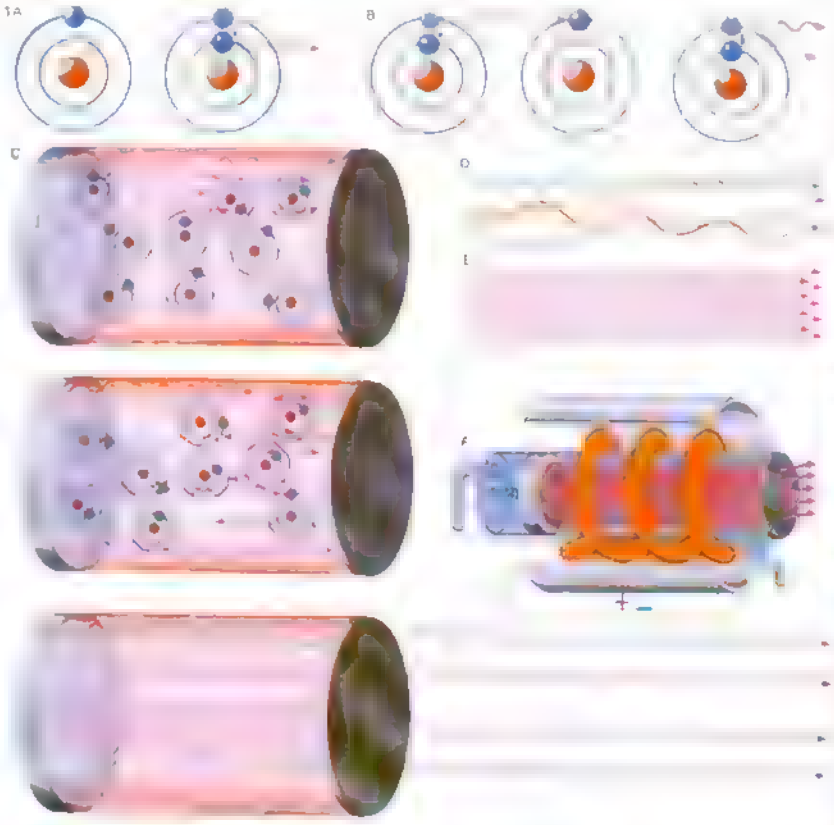
Each photon triggers the production of another one and so they all travel together and produce light waves that are exactly in step. This light is said to be in phase, or coherent. (In ordinary light, the waves are all out of phase.) Because the waves are all in step, they reinforce each other and laser light is very bright. The construction of the laser produces a narrow beam that hardly spreads at all – even at the distance of the Moon, a laser beam directed from the Earth is only 3km (2 miles) wide [3]. A narrow beam of intense coherent light is extremely concentrated in energy and, if a laser beam is focused to a point by a lens, it will heat the air to a state of incandescence (bright and glowing with heat) or burn a hole in a steel plate. A straight narrow beam of laser light can be used for precise alignment in the construction of tunnels and pipelines, for example. The beam is directed along the proposed route and can be seen by the construction engineers only when they are directly in line with it.

Other uses of lasers

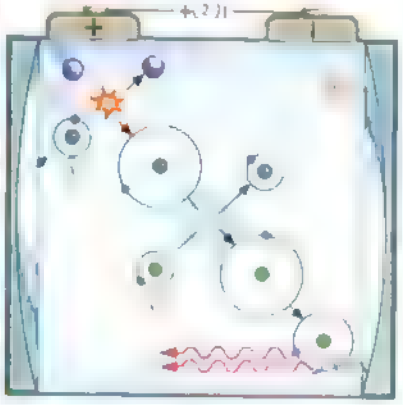
Laser beams can also be used to measure distances and speeds. These have included firing a laser beam at the moon to reflect it from a

CONNECTIONS

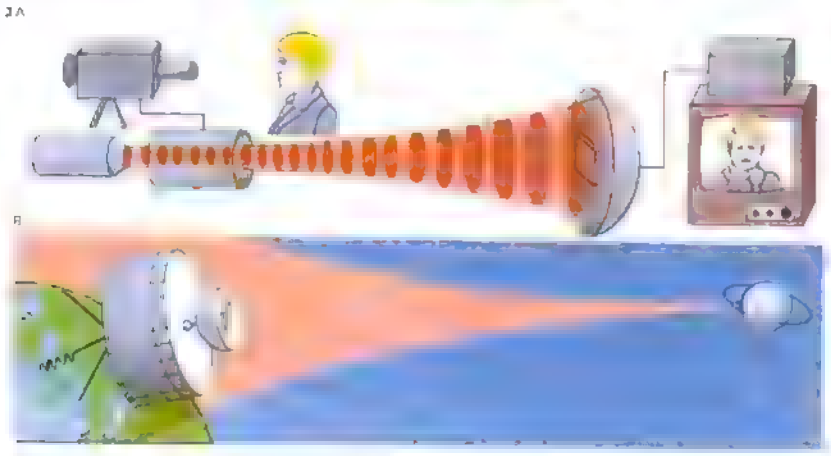
See also



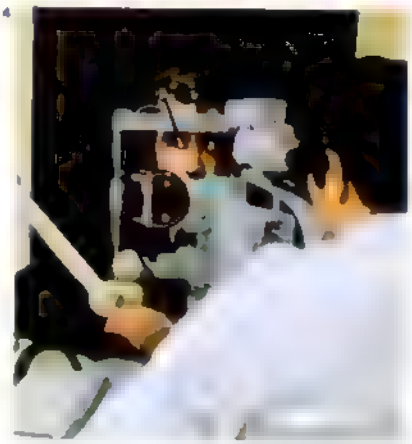
1 Normal emission of light occurs when an electron in a high energy orbit falls to low orbit [A]. Stimulated emission [B] is triggered by light emitted from another atom. In a laser [C] most atoms are brought to a high energy state by pumping in energy. Some begin to produce light by normal emission and mirrors at each end reflect the light to and fro, producing stimulated emission until all the atoms are in a low energy state. The light leaves the laser through one of the mirrors. Ordinary light [D] is a mixture of different frequencies moving in various directions, whereas laser light [E] has a single frequency and moves in the same direction with all waves in phase. The first laser [F] contained a synthetic ruby crystal surrounded by a flash tube (to pump in light energy) and a pair of reflecting mirrors.



2 The helium-neon gas laser contains two gases and operates continuously. Ions [1] accelerate in an electric field produced by a high voltage [2] to excite helium atoms [3]. Through collisions they transfer their energy to equal energy levels in the neon atoms [4]. Light amplification occurs when a passing photon stimulates excited neon atoms to radiate. After several reflections from partly silvered mirrors [5] the laser beam emerges.



3 Laser communication would be ideal for interplanetary missions because the narrow, powerful beam [B] can reach a small, distant target. A laser carries a television or other signal by modulating the beam at the transmitting end, focusing the beam on a detector at the receiving end and demodulating the signal produced. [A] Ordinary light is unsuitable for communication because it has many frequencies that interfere with each other, only the laser can be used.



4 A detached retina is quickly and painlessly welded into place by a laser beam, instantly curing partial blindness. The surgeon lines up the laser by directing a beam of ordinary white light into the unaffected eye. When the white beam is in place the laser is fired briefly and the green laser light enters the other eye and is focused on the affected retina to seal it in place.

special mirror placed there by the Apollo astronauts and thus give a very accurate measure of the moon's distance.

In meteorology, laser beams are used to detect invisible air layers and movements as well as clouds, and they are useful in studies of air pollution.

The intense heat of lasers gives them all kinds of uses in medicine and industry. A laser beam directed into the eye at insufficient power to damage the lens is focused by the lens onto the retina, where it can painlessly weld a detached portion back in place and restore failing sight [4]. Laser beams can burn away skin growths without surgery, by firing the beams along fibre-optic tubes inserted into the body, and painlessly drill decayed teeth. In industry, lasers cut out patterns, drill holes in diamonds to make dies for wire manufacture and shape and weld parts for microelectronic circuits [Key].

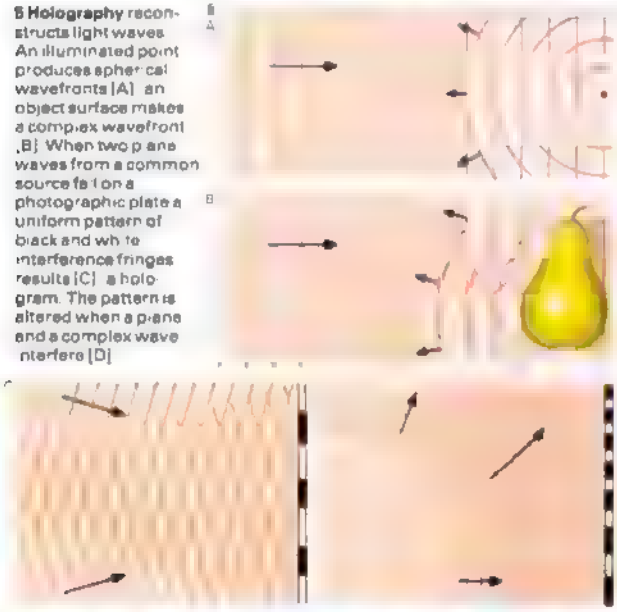
Communication by laser beams instead of radio waves is desirable because light beams can carry many more channels of information than can radio. Data, sound and pictures can be transmitted by a laser beam, which routes it along an enclosed path of some kind to avoid loss of signal strength from having to pass through fog or mist in the air.

One of the most amazing consequences of producing coherent light in lasers is the development of holography, with which three-dimensional images can be made [5, 6, 7]. Although three-dimensional colour television and motion pictures may one day result from it, holography has several uses now. Double-exposure holograms record any movement of the object between the exposures and so readily picture the vibrations in a surface. Vibration analysis is essential to the design of components such as aircraft and engine parts which must perform faultlessly for long periods at high speeds and stresses.

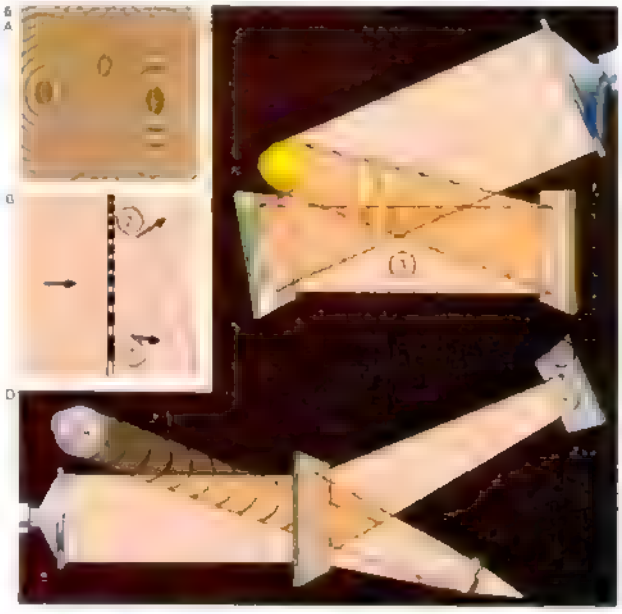
Another field that may be revolutionized by the laser is nuclear energy. Research is being carried out to see if thermonuclear fusion (the reaction that takes place in a hydrogen bomb and in the stars) can be initiated by a laser beam instead of producing a high-temperature plasma by means of a powerful electric discharge.

KEY

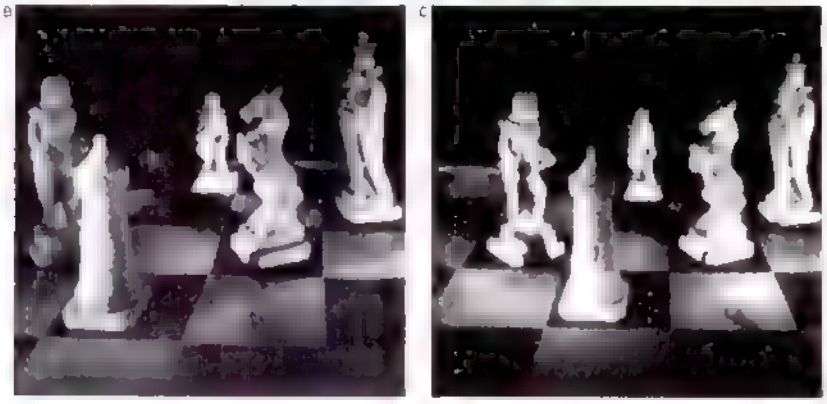
A laser beam contains sufficient energy to "burn" a hole in hard materials such as steel and diamond. Here a laser beam is drilling a hole in a sheet of toughened glass. Laser-drilled diamonds are used as dies for drawing metal into extremely fine wires.



5 Holography reconstructs light waves. An illuminated point produces spherical wavefronts [A]; an object surface makes a complex wavefront [B]. When two plane waves from a common source fall on a photographic plate a uniform pattern of black and white interference fringes results [C], a hologram. The pattern is altered when a plane and a complex wave interfere [D].



6 When a laser beam passes through a hologram of circular fringes [A] two wave fronts emerge [B]: one converges [1] to form a real image, the other diverges [2] as if from a virtual image. A hologram is made [C] by illuminating an object by a laser light, which then falls on a photographic plate. Part of the laser beam is reflected directly on to the plate as a reference beam [3]. The image is recreated [D] by illuminating the hologram with laser light. An observer sees a three-dimensional virtual image of the object [4], the appearance of which changes with viewing angle. A real image can be recorded photographically [5].



7 Images in 3-D can be obtained by holography. A hologram of an object is made by the technique described in illustration 6, using apparatus such as that shown in A. It is arranged so that laser light reflected from the object travels about the same distance as the reference laser beam reflected onto the photographic plate by a mirror. When the hologram is illuminated from behind with laser light, an image of the object is seen which is not only 3-dimensional but also demonstrates parallax: as the head is moved, a different view of the object is obtained, as in real life. This is evident in two views of chessmen [B, C]. The image has depth because the hologram completely reconstructs the light rays coming from the object and striking the plate. The colour of the object can also be produced by using several lasers of different colours. It is possible that holography may one day give us a totally realistic cinema or television picture.

Scientific and astronomical tables

SYMBOLS USED IN MATHS AND SCIENCE

Symbol	Meaning
+	Plus, positive
−	Minus, negative
×	Multiplied by
÷	Divided by
±	Plus or minus
∓	Minus or plus
:	Ratio, in proportion to
=	Equals
≠	Does not equal
≡	Is identical with
≢	Is not identical with
≈	Approximately equals
~	Equivalent or similar to
>	Is greater than
≯	Is not greater than
<	Is less than
≮	Is not less than
≥	Is greater than or equal to
≤	Is less than or equal to
∝	Is directly proportional to
→	Approaches as a limit
x^n	x to the nth power
$\sqrt{x}$, $x^{1/2}$	Square root of x
$\sqrt[n]{x}$, $x^{1/n}$	The nth root of x
x^{-n}	Negative exponent
$n!$ or $\underline{|n}$	Factorial of x (Here $1 \times 2 \times 3 \ldots x$)
∞	Infinity
f, F, ϕ	A function of
∠	(In geometry) an angle
⊥	Perpendicular, is perpendicular to
∥	Parallel, is parallel to
△	Triangle
≅	Congruent, is congruent to
∴	Therefore
∵	Because
π	Pi, 3.14159
$x°$	Degree of angle [x degrees]
x'	Minute of angle [x minutes]
x''	Second of angle [x seconds]
sin	Sine
cos	Cosine
tan	Tangent
ctn, cot	Cotangent
sec	Secant
cosec	Cosecant
covers	Conversed sine or coversine
exsec	Exsecant
e	Base of the natural logarithms
$\log_a x$	Logarithm (base a) of x
$\ln_e$, $\log_e$, $\log_n$	Natural logarithm of a
$\log_{10} \log a$	Common logarithm of a
antilog	Antilogarithm
colog	Cologarithm
exp x	e^x
GCD, gcd	Greatest common divisor
LCD, lcd	Least common denominator
LCM, lcm	Least common multiple
Σ	Sum of specified terms
∫	Integral
$\lim_{x \to a} y = b$, $\lim_{x \to a} y = b$	The limit of y as x nears a is b
Δy	An increment [of y]
δy	A variation or increment [in y]
dy	Differential of y
∋	(In logic, such that
~p, p̄, p′, p	Not p
p ∧ q, p.q, p&q	Both p and q
p^2 q	At least one of p and q
p/q, p|q	Not both p and q
p↓q, p ′ q	Neither p nor q
p→q, p⇒q, p⊂q	If p then q, only p if q
V, 1	Universal class
Ø, ∧, 0	Null class
a∈ M	The point a belongs to the set M
M = N	The sets M and N coincide
M ⊂ N	M is a subset of N
M ⊆ N	M is a subset of N
M ∩ N, MN	Intersection of M and N
M + N, M ∪ N	The sum of M and N
M − N, M ∖ N	All points of M not in N
~M, C(M), M′	The complement of M

BASE SI UNITS

Physical quantity	SI unit	Symbol
length	metre	m
mass	kilogram (me)	kg
time	second	s
electric current	ampere	A
thermodynamic temperature	kelvin	K
luminous intensity	candela	cd
amount of substance	mole	mol

SUPPLEMENTARY SI UNITS

plane angle	radian	rad
solid angle	steradian	sr

DERIVED SI UNITS WITH SPECIAL NAMES

Physical quantity	SI unit	Symbol	Physical quantity	SI unit	Symbol
frequency	hertz	Hz	magnetic flux	weber	Wb
energy	joule	J	inductance	henry	H
force	newton	N	magnetic flux density	tesla	T
power	watt	W	luminous flux	lumen	lm
pressure	pascal	Pa	illuminance	lux	lx
electric charge	coulomb	C	activity (radioactive)	becquerel	Bq
potential difference	volt	V	absorbed dose	gray	Gy
electric resistance	ohm	Ω			
electric conductance	siemens	S			
electric capacitance	farad	F			

SI PREFIXES (DECIMAL MULTIPLES AND SUBMULTIPLES)

Submultiple	Prefix	Symbol	Submultiple	Prefix	Symbol	Multiple	Prefix	Symbol	Multiple	Prefix	Symbol
10^{-1}	deci	d	10^{-9}	nano	n	10^{1}	deca	da	10^{9}	giga	G
10^{-2}	centi	c	10^{-12}	pico	p	10^{2}	hecto	h	10^{12}	tera	T
10^{-3}	milli	m	10^{-15}	femto	f	10^{3}	kilo	k	10^{15}	peta	P
10^{-6}	micro	μ	10^{-18}	atto	a	10^{6}	mega	M	10^{18}	exa	E

IMPERIAL UNITS OF LENGTH

4 inches	1 hand(breadth)
3 barleycorns	1 inch
12 inches	1 foot
3 feet	1 yard
5.5 yards	1 rod, pole or perch
4 rods	1 chain
10 chains	1 furlong
8 furlongs	1 mile
3 miles	1 league
6 feet	1 fathom
2.5 fathoms	1 nautical chain
400 nautical chains	1 nautical mile

IMPERIAL UNITS OF CAPACITY

60 minims	1 fluid drachm
8 fluid drachms	1 fluid ounce
5 fluid ounces	1 gill
4 gills	1 pint
2 pints	1 quart
2 quarts	1 pottle
2 pottles	1 gallon
2 gallons	1 peck
4 pecks	1 bushel
4 bushels	1 coomb
2 coombs	1 quarter

UNITS OF MASS – AVOIRDUPOIS

16 drachms	1 ounce
16 ounces	1 pound
14 pounds	1 stone
2 stones	1 quarter
4 quarters	1 hundredweight
20 hundredweights	1 ton
7000 grains	1 pound

UNITS OF MASS – TROY

4 grains	1 carat
6 carats	1 pennyweight
20 pennyweights	1 ounce
12 ounces	1 pound (troy)
25 pounds (troy)	1 quarter
4 quarters	1 hundredweight
20 hundredweights	1 ton

The hundredweight and ton are sometimes called the long hundredweight and long ton to distinguish them from the US units, the short hundredweight (100 pounds) and the short ton (2,000 pounds).

UNITS OF MASS – APOTHECARIES'

20 grains	1 scruple
3 scruples	1 drachm
8 drachms	1 ounce
12 ounces	1 pound

INTERCONVERSION FACTORS

Length	metres	inches	yards
1 metre	1	39.3701	1.0938
1 inch	0.0254	1	0.2778
1 yard	0.9144	36	1

	kilometres	miles	nautical miles
1 kilometre	1	0.62137	0.53996
1 mile	1.60934	1	0.86898
1 nautical mile	1.852	1.1508	1

1 light-year	9.4607×10^{15} metres
1 parsec	3.0857×10^{16} metres
1 astronomical unit	1.495×10^{11} metres

Area	sq metres	sq inches	sq yards
1 sq metre	1	1.550	1.196
1 sq inch	0.000645	1	0.00077
1 sq yard	0.8361	1,296	1

	sq kilometres	sq miles	acres
1 sq kilometre	1	0.386	247.105
1 sq mile	2.58999	1	640
1 acre	0.00405	0.00156	1

1 hectare = 0.001 sq km = 2.47105 acres

Volume	cubic metres	cubic feet	gallons
1 cubic metre	1	35.3146	219.969
1 cubic foot	0.0283	1	6.2288
1 gallon (UK)	0.00455	0.160544	1

1 gallon (US) = 0.83268 gallon (UK)

Mass	kilograms	pounds	tons
1 kilogram	1	2.2046	0.000984
1 tonne	1,000	2,204.6	0.98420
1 pound	0.45359	1	0.000446
1 ton (long)	1,016.047	2,240	1

Energy	joules	cals	K Wh	btus
1 joule	1	0.2388	2.778×10^{-7}	9.478×10^{-4}
1 calorie	4.1868	1	1.163×10^{-6}	0.00397
1 kilowatt hour	3.6×10^{6}	8.598×10^{5}	1	3,412.1
1 British Thermal unit	1,055.06	251.997	2.931×10^{-4}	1

Forked flashes

What is electricity?

To the man in the street, electricity is the cause of a lightning flash [1–3] or the form of energy that powers his television set and washing machine. He knows that electric trains use electrical power and he is reminded of his dependence on it by the network of power lines criss-crossing the countryside, or by a power cut, when he has to read by candlelight. But there are other less well-known everyday processes that involve the use of electricity. A beating heart, a running athlete, a dreaming baby and a swimming fish all generate a form of electricity just as surely as a power station does.

Electrons and protons

To a scientist, electricity results from the movement of electrons and other charged particles in various materials. A scientific understanding of electricity therefore depends on a knowledge of atoms and the sub-atomic particles of which they are composed. The key to this understanding is the tiny electron – tiny even when compared with the minute atom in which it may be found.

Atoms of all materials have one or more electrons circling in orbits of various sizes – much as the planets move round the sun. Normally the number of electrons equals the number of protons in the nucleus. The protons, however, being much heavier than the electrons, are virtually stationary in the atom's centre. This extremely simplified model of the atom is sufficient to explain the basis of electricity.

The electrons and protons each have an electric charge (but of opposite polarity) and attract each other. Charges of the same polarity repel each other. To distinguish the proton's charge from that of the electron the former is called positive and the latter negative. An atom that has more, or fewer, electrons than normal is called an ion. If it is deficient in electrons, it is called a positive ion; if it has an excess of electrons, it is called a negative ion.

When an electron moves away from an atom the atom is left with a net positive charge. The electron, deprived of its positive counterpart in the atom's centre, eventually moves about another atom or possibly returns to the ion it has left.

Why do electrons move at all? There are a number of possible causes. A common one is simply that, if an incoming electron or light pulse hits an atomic electron, the latter can be knocked out of its orbit. Heat makes atoms dance faster, causing the electrons to move so energetically that they may shoot away from the parent atoms. Chemical activity will also cause electrons to move out of atoms.

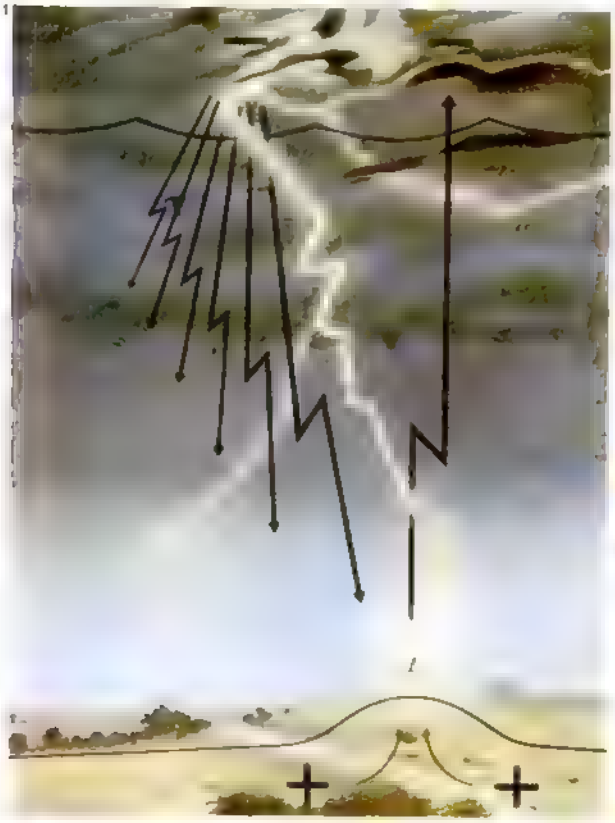
A good example of the relationship between chemical and electrical activity is found in the muscles. Muscle fibres contract when they are electrically stimulated [4]. Normally, this is caused by the release of a chemical from an associated nerve, following the receipt of an electrical signal from the nervous system. When part of this system is damaged and muscles become weak or fibres are destroyed, it is possible to apply external electrical signals to stimulate muscle activity and strengthen their fibres.

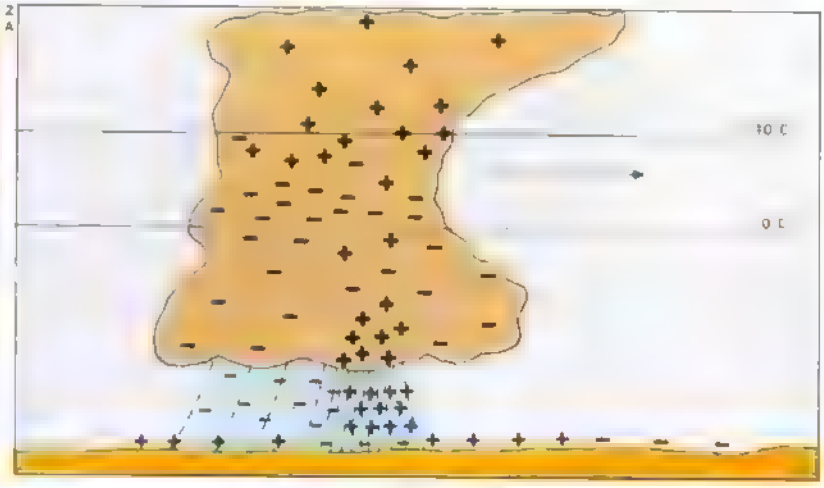
Conductivity

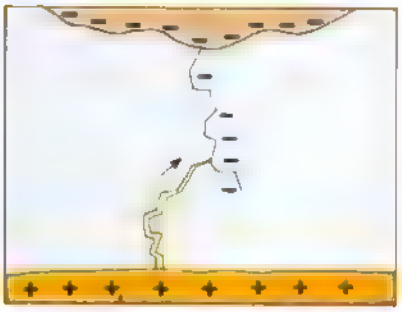
The electrons of some materials move more freely than others. This characteristic is known as conductivity. Most metals, hot

CONNECTIONS

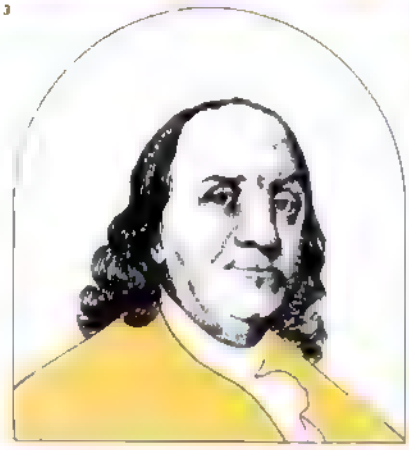
See also



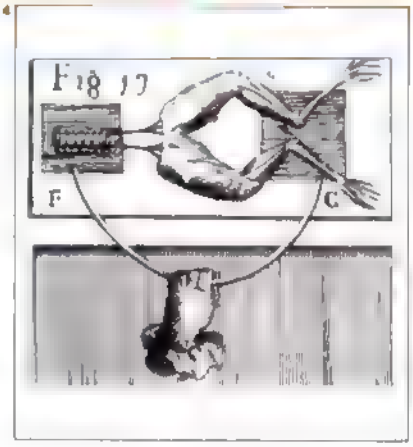




1 Lightning strokes were not properly explained until about 200 years ago. They are caused by an electrical imbalance between clouds or between clouds and the earth. The base of the cloud can have an excess of electrons and then draws positive ions to the ground underneath it. The potential difference grows until there is a sudden flow of electrons (the flash), neutralizing charges on both the ground and the cloud.



3 Benjamin Franklin (1706–90) was the first man to recognize the true nature of lightning. During a thunderstorm he induced a flash of lightning to flow along the string of a kite to the earth.



4 Galvanism was the term used to describe the twitching effect produced by an electric current on a pair of frog's legs. Luigi Galvani (1737–98) used this to show the connection between muscle activity and electricity.

2 Formation of lightning starts with a big storm cloud (A), within which there is a significant temperature difference. Electrons move downwards and positive ions move upwards within the cloud, causing positive ions to gather on the earth below. When there are sufficient electrons, a sudden breakdown of the air occurs and a stream of electrons shoots earthwards (B) to be met by an upward stream of ions (C).

gases and some liquids are good conductors. Air, rubber, oil, polythene and glass are bad conductors so that they can be used to cover good conductors without themselves taking part in electron flow [6].

These bad conductors are called insulators. No insulator is "perfect". Under certain circumstances the electrons of any atom can be forced out of it. But the conditions required are generally so unusual and difficult to arrange with these materials that, for practical purposes, they can be considered inactive.

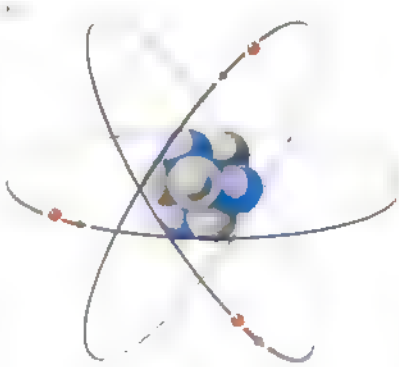
There is also a group of materials – the semiconductors – that behave partly as insulators and partly as conductors. Among these are germanium, silicon and copper oxide. Their properties can be exploited for many purposes. For example, using one of the semiconductors it is possible to make an electric "valve" that, like the valve on a bicycle tyre, allows easy electron movement in one direction only. This device is called a rectifier, it is used in both tiny radio sets and large power stations to change an alternating current to a direct current.

Heat is simply a chaotic form of molecular activity or electronic motion and temperature is a measure of its vigour. When the temperature of most metals is reduced, it is easier for electrons to move freely, that is, the electrical resistance (to free electron movement) falls as the temperature drops and the conductivity of the metal increases.

Superconductivity

If in certain materials the temperature drops low enough, the resistance to electron flow ceases completely and electrons, once started on their journey, continue to move indefinitely provided the temperature is kept sufficiently low. The condition of zero resistance is called superconductivity. It occurs in metals such as tin, lead, aluminium and niobium [7] at a few degrees above absolute zero (−273°C or −460°F).

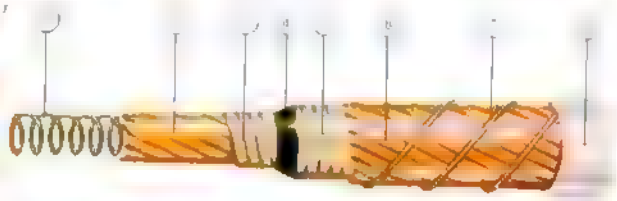
Electricity, therefore, is simply the movement of electrons or other charged particles. These particles are among the smallest components of matter, and yet the way in which they move and interact has a great influence on every aspect of life.



The electron's charge is the basic unit of electricity. The electron is a fundamental particle found in all kinds of atoms. In this simple model of an atom of the metallic element lithium, three electrons (red) can be seen circling the central nucleus. Protons (blue) and neutrons (grey) make up the nucleus. Each electron carries a negative electric charge and each proton carries a positive one, so that the three electron charges are exactly balanced by the three proton charges, making the whole atom electrically neutral. In a conductor – as in most metals – an external electromotive force (voltage) causes electrons to drift from atom to atom and it is this flow of electrons that constitutes an electric current. Electron movement occurs because in a conductor the outermost electrons are not tightly bound to its nucleus. In a non-conductor or insulator the electrons are too tightly bound to leave the nucleus easily and so such substances do not conduct electricity. In some situations atoms can completely lose or gain one or more electrons to become permanently charged. Such charged atoms, called ions, can also act as current carriers.

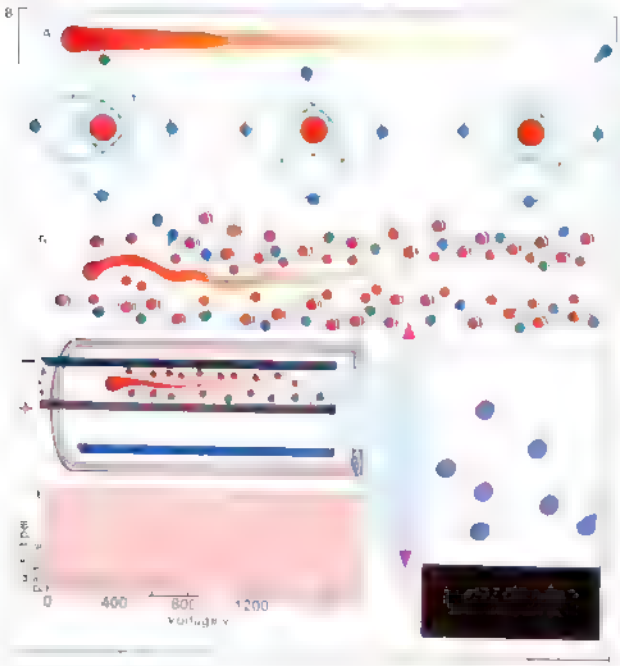
5 The beating heart generates tiny electrical currents which, after being suitably amplified, can be displayed on a cathode ray tube. These current shapes can be recorded permanently on paper, in which form they are called electrocardiograms for the heart [1] and electroencephalograms [2] for the brain. Existing or potential malfunctions of both these vital organs can be diagnosed with their aid.

6 Domestic wiring uses copper wires encapsulated in rubber or plastic [A, C]. In fire-proof wiring [B], the wires are embedded in a non-inflammable powder surrounded by a copper tube.

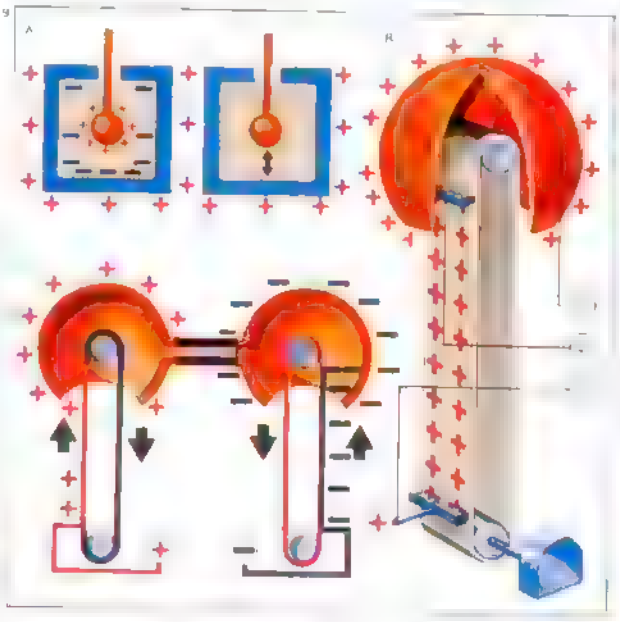


1 Flexible conductor
2 Strips of inner conductor
3 Inner conductor screen
4 Lapped tape dielectric
5 Outer conductor screen
6 Strips of outer conductor
7 Helical skid wires
8 Hard bore pipe

7 Perfect conductors can be made out of alloys of metals (eg tin, lead and niobium) at temperatures close to absolute zero (−273°C). Once electrons become detached from their parent atoms they move through the supercooled conductor, without slowing down or coming to rest, for indefinitely long periods. At such low temperatures the atoms in the metal vibrate only very slightly.



8 A high velocity particle passing through a gas knocks electrons off the otherwise neutral atoms [A]. As it sweeps through the gas, it leaves behind a stream of free electrons [B] shown here in blue. If these break free with sufficient energy, they can knock other electrons out in a Geiger counter [C, D], subatomic particles enter the chamber where they are accelerated to produce more free electrons. These are attracted to the positive plate and are led from there to drive a meter or earphones. The current flow depends on the voltage v between the plates [E]. The meter reading or frequency of clicks indicates the amount of radioactivity of the source.



9 Extremely high voltages can be produced with the Van de Graaff generator [B]. If a body having an excess of positive ions is placed inside a container, the inside acquires electrons [A] and the outside an equal number of positive ions. If the charged body touches the inside all the free electrons flow into it, thus making it neutral. The outside of the container still retains its positive ions. In the Van de Graaff generator, positive ions are sprayed from a suitable source [1] onto an endless conveyor belt which carries them inside a metal sphere. The belt connects to the inside wall through a conductor in the form of a comb [2], thus permitting an electron flow to the belt. This causes positive ions to form on the sphere's outside wall [3]. The effect may be enhanced by using two generators connected as shown [C].

What is an electric current?

Electricity flowing along a wire is known as an electric current. The wire is the conductor. When an electric lamp is connected across a battery and switched on, current flows along a wire from one terminal of the battery to the lamp, through its filament, making it glow white hot, and back along a second wire to the other battery terminal. If the switch is turned off, the circuit is broken, current flow stops and the lamp is extinguished.

Movement of electrons

The current carriers in most circuits are electrons from the metal making up the conductors. In all conductors, and a few other materials as well, there is always a random movement of electrons (minute charged particles), even when no current flows. The electrons may be relatively free to move or more tightly bound. Good conductors have freer electrons and hence more electron movement than do bad conductors, or insulators, in which most of the electrons are too tightly bound to their parent atoms to move easily. Sometimes, through natural or contrived processes, there can be a net movement of electrons in a specific direction. This concerted flow is the electric current and it is measured in amperes, generally given by the abbreviation A. Other current carriers include ions (charged atoms or molecular fragments) in gases and solutions, and "holes" (a deficiency in electrons in some types of semiconductors – the holes behave as positively-charged carriers of electric current).

A force has to be applied to cause a net flow of electrons in one direction. In nature this force can be derived from a number of sources such as sunlight, magnetic action or chemical activity. Some of these have been exploited to generate electric current. Two common devices designed for this purpose are the generator [9], which utilizes magnetic effects, and the cell (sometimes called a battery) [6], which depends on chemical action. Both devices force electrons to move in one direction round a circuit by virtue of the electromotive force (emf) they generate. The electromotive force is measured in volts, using a voltmeter.

The voltage of an emf and current flow are related in a similar way to water pressure and water flow. In a household, all the pipes are full of water at a certain pressure. But there is no movement until a tap is opened, allowing water to run.

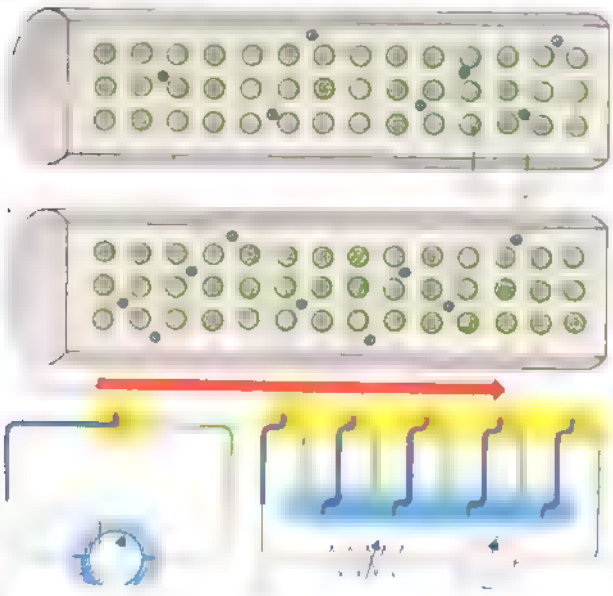
An electrical circuit may be connected to a source of emf without causing any specific electron flow (current) until a path is provided through which the electrons can move. This may be a light bulb [Key] or a vacuum cleaner; an electric switch is like a tap that turns on the current.

Relationship between voltage and current

As the voltage in a circuit increases, so does the current. An electrical circuit is, however, made up of a number of different parts. There is normally a switch, conductors and the appliance that is being supplied with electricity. All these taken together have a resistance to current flow, which is constant (provided the temperature remains the same) for that particular group of components. Therefore although the same voltage may be applied to a light bulb and an electric iron, the actual current flow is different in each, because each has a different resistance. So it

CONNECTIONS

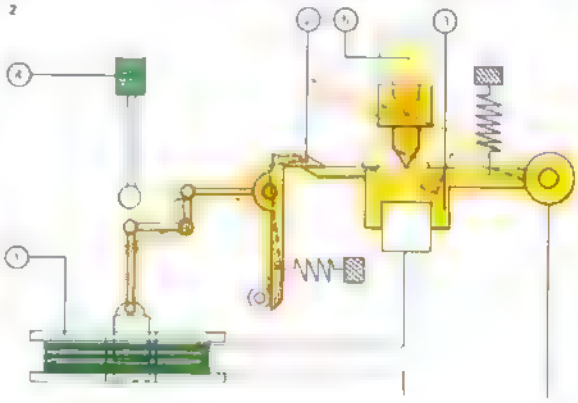
See also
What is electricity?
Magnets and
Basic electronic principle
Transformer a

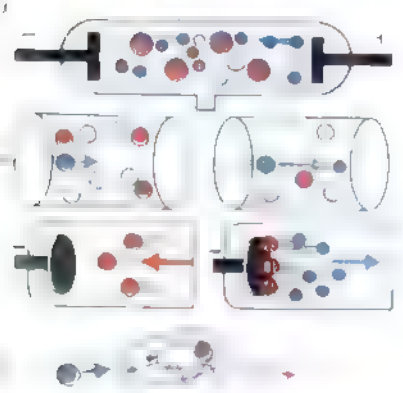


1 Metal is a good conductor of electricity because its atoms [1] easily give up their outer electrons [2], which move freely within the crystal lattice of the metal [A]. An electric current flows only if the "free" electrons have a net motion in one direction [B]. A voltage [3] from a battery, say, can cause such a flow. In a thermocouple [C], a temperature difference between the two junctions of a circuit, made of two dissimilar metals, will produce a current, indicated by a meter [4], which is proportional to the temperature difference. In a thermopile [D], several thermocouples are connected in series to increase the voltage of the circuit.

2 Circuit-breakers can be used either in conjunction with, or in place of, a fuse to interrupt dangerously high currents. When too large a current flows through the coil [1], a magnetic field is produced which activates the catch [2], causing the spring-loaded contact [3] to rise. The current is interrupted, thereby protecting the circuit of which the breaker is a part. No further current can flow until the circuit breaker is closed again, which is achieved by pressing the reset button [5]. To test the circuit-breaker, or to operate it manually, a push button [4] is pressed to move the contact.

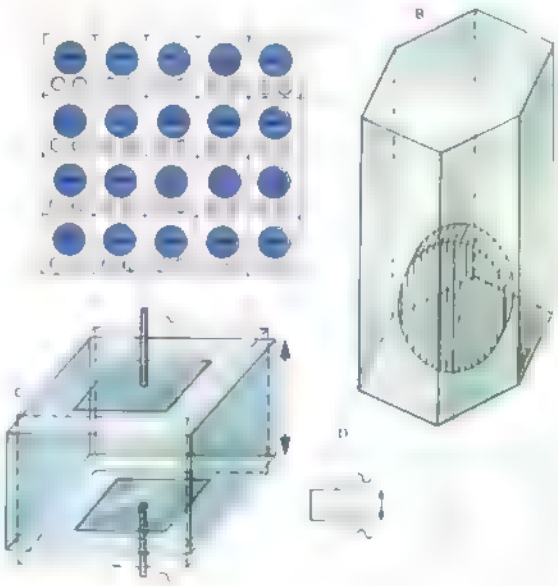
2





3 A gas discharge lamp gets its light from the energy changes in gas atoms. A positive and negative electrode (anode and cathode) at opposite ends of a gas-filled glass tube [A] attract electrons and positive ions. **Reducing the pressure** [B] speeds this up. As ions hit the cathode [C] they dislodge electrons which speed towards the anode, colliding with gas atoms on the way. The atoms absorb the energy for a moment, then release it in the form of light [D].

4



4 Certain crystals, including quartz and sapphire, exist in the form of "cells" in a delicate state of electrical equilibrium [A]. Applying a varying voltage across the crystal disturbs the delicate balance and causes the crystal to vibrate and emit sound or ultrasonic waves, generally at a specific angle to the direction of the applied voltage [B]. Conversely, when such a crystal is vibrated, it generates a voltage. This effect – which is known as piezoelectricity – is made use of in a gramophone pick-up [C] and in crystal microphones. In the pick-up, the movement of the stylus in the record groove rapidly vibrates a piezoelectric crystal and generates a small electric current. In a crystal microphone, sound waves vibrate a diaphragm coupled to a crystal and generates a current that can be amplified and fed into a tape recorder or sound system. Most crystals respond strongly to only one frequency, depending on their dimensions. Radio transmitters make use of this property to "hold" a particular fixed frequency [D]. Vibrating quartz crystals are able to keep almost perfect time. They are used in quartz clocks and watches, which are accurate to within a few seconds over a period of several years. Piezoelectric crystals are also used to generate the electricity for the spark that ignites the gas in certain kinds of "electronic" cigarette lighters, which do not need any batteries.

is not only the magnitude of the voltage that determines how much current flows through a particular piece of equipment, it is also the resistance of that equipment and the conductors. This property of electrical resistance is measured in ohms (Ω). For any conductor, or system of conductors and equipment, the relationship between voltage, current and resistance is given by the formula: voltage = current × resistance. This is the mathematical expression of Ohm's law, named after Georg Ohm (1787–1854) who was the first person to specify the interdependence of these three factors in a precise way.

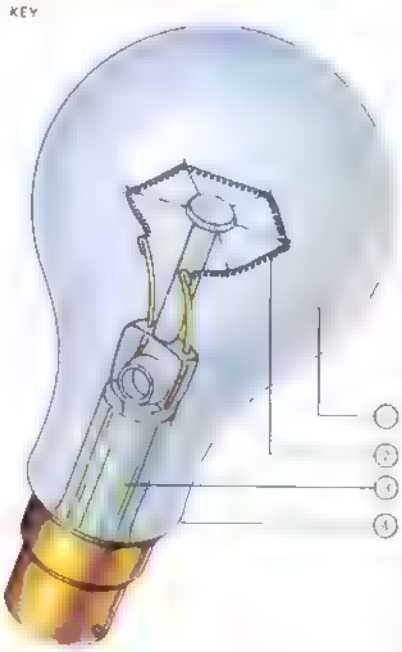
The resistance of electrical conductors depends on their dimensions and on the materials they are made of. As the cross-sectional area increases, the resistance falls, but as the length increases, the resistance rises. A long thin conductor therefore has more resistance than a short thick one with the same volume of material. Silver has a lower resistance than copper, whereas aluminium and iron have higher resistances.

Current flows at the same rate at all points around a circuit at any one time. According to a convention adopted before the nature of electricity was properly understood, a direct current (that is, current from a battery or dynamo) is assumed to flow from a positive point to a negative one. As it happens, electrons move from negative to positive, so that electron movement is opposite to the assumed current flow.

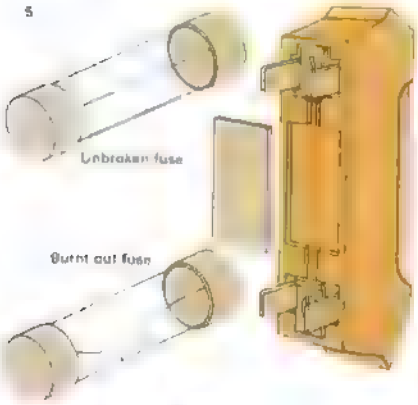
Effects of current flow

Three phenomena that typically occur when a current flows (and by which it can be detected) are heating, chemical and magnetic effects. Its heating effect is used to provide warmth in electric fires, cookers and industrial furnaces. Such heating can also be unwanted; large cables carrying thousands of amperes have to be cooled to prevent the current-generated heat melting the insulation or even the wires themselves.

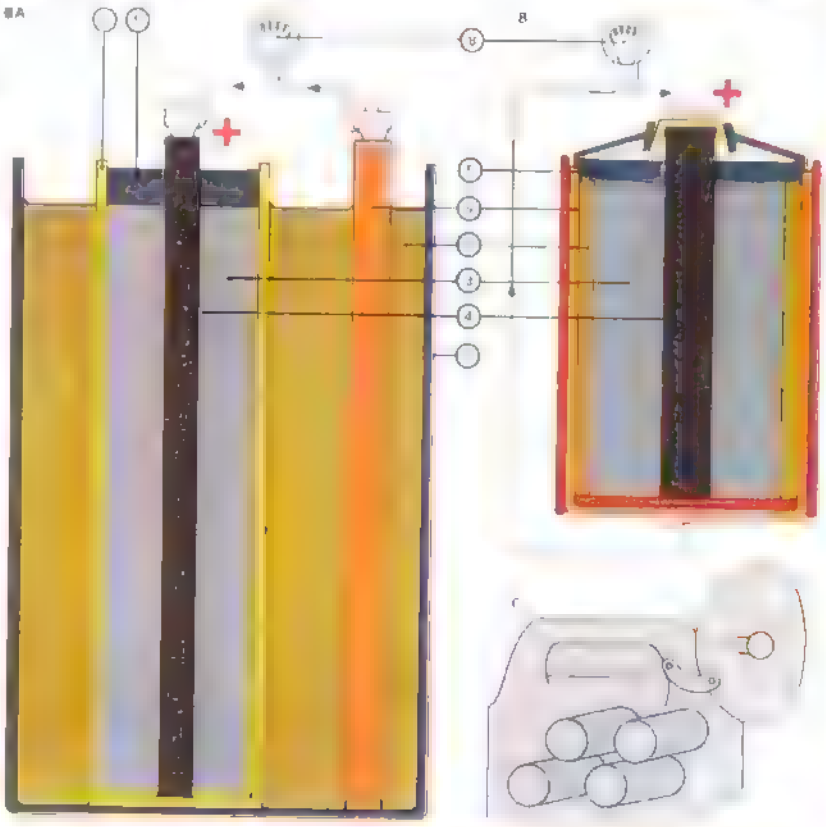
The chemical effect of current is used in electroplating and in energy storage, particularly in cells, the most familiar of which is the lead-acid accumulator or battery [7]. The magnetic effect is used in motors, electromagnets and many other devices.



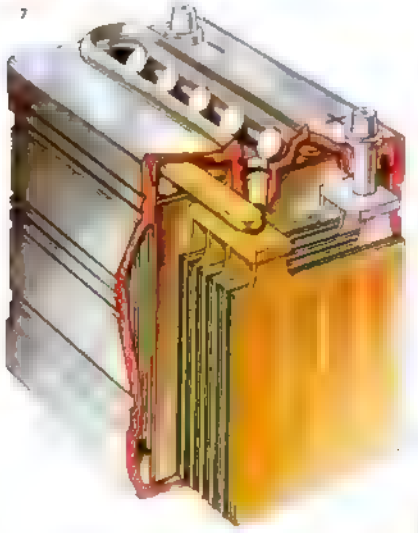
Heat generated by the passage of electric current is the source of light in a so-called "filament" or "incandescent" lamp. Because it is enclosed in a vacuum, or inert gas [1], the filament [2] cannot oxidize when the current passes through, causing it to become hot. It is made of a tungsten alloy, combining mechanical and thermal strength. Though it is extremely thin, it glows white hot when enough current passes through it. It is supported on two glass columns [3] through which the connecting wires pass. The whole assembly is enclosed in a thin glass envelope [4]. Only about two per cent of the electrical energy is converted into light.



5 The heating effect of an electric current is used in a fuse, which consists of a thin wire that melts when excessive current passes through it, thereby cutting off the electricity supply.

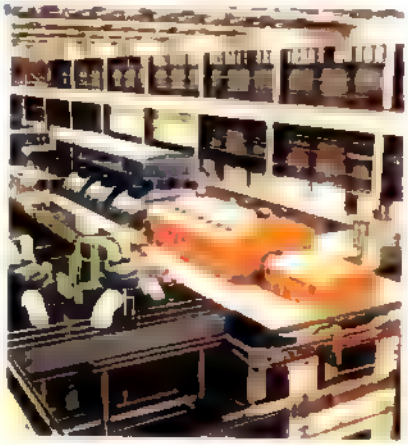


6 The Leclanché cell [A] consists of a leakproof jacket [1] containing a porous pot [2] in which there is a paste of manganese dioxide and carbon granules [3] surrounding a carbon rod [4]. The top can be sealed with pitch [5]. A zinc rod [6] stands in a solution of ammonium chloride [7], and is connected to the carbon rod via a circuit and a light bulb [8]. The zinc dissolves in the solution, setting up an electromotive force. The ammonium ions migrate to the carbon anode and form ammonia (which dissolves in the water), and hydrogen ions. Torch dry batteries [B and C] use wet paste cells of the Leclanché type.



7 A 12-volt car battery has six two-volt cells connected in series. The cells have anodes of brown lead oxide and cathodes of porous grey lead immersed in sulphuric acid. An electric current flows if the electrodes are connected through a conductor. When the battery supplies current the sulphuric acid converts the anode to lead sulphate, thus reducing the strength of the acid. This process is reversed during recharging. Each cell of the battery is made of several anodes and cathodes separated by porous insulators. The cells are housed in a hard rubber case and the various cells are interconnected with lead bars.

8 Tiny electric motors, such as this one, which is used for driving a miniature tape recorder, emphasize the enormous range of size and applications of electrical equipment.



9 The heart of a power station is the generator, which sends electricity over hundreds of kilometres through a wide network of transmission lines. It is the point at which mechanical energy is converted to electrical energy.

Magnets and magnetism

Magnetism and electricity are not two separate phenomena. The error of thinking they were arose from the fact that their interrelation was not appreciated until 1820. In that year the Danish scientist Hans Christian Oersted (1777–1851) showed that an electric current flowing in a wire deflects a compass needle close by. Whenever an electric current flows, whether from cloud to ground in the form of lightning or through a muscle in the body, a magnetic field is created.

Thousands of years before electricity was recognized and used, magnetism was observed and applied – mainly for navigation. Eventually, when science became aware of the atomic nature of matter, it was finally realized that the properties of magnetism and electricity are both bound up in the nature of the physical structure and arrangement of atoms and their electrons.

Whenever magnetism can be detected there must be a current of electricity. Those materials that appear to be magnetic without any external source of electricity depend on electron movements within their atomic structure to provide the electric current; this is the class of magnetism dealt with here.

The property of attracting iron and iron-based materials occurs naturally in a mineral called lodestone [3], itself a chemical compound of iron. It is likely that some form of lodestone was used in the first magnetic compasses that the Chinese are thought to have made [1]. It is relatively easy to transfer deliberately magnetic properties between various materials, of which iron and steel are the best known common examples.

Permanent magnets

Iron-attracting materials form a class of so-called permanent magnets, although they may retain their magnetic properties for only a limited time. In the form of a bar, a permanent magnet experiences a force due to the earth's magnetism such that, if it were free to move, one end would point roughly in the direction of the earth's North Pole and the other to the South. The two ends are named the north-seeking (or north) pole and south-seeking (or south) pole.

Unlike magnetic poles attract each other. A magnet that attracts other material does so by first turning the material into a weak magnet. Like poles repel each other (although this is not so obvious as attraction) because whenever an iron or steel object comes within the influence of a magnet, and itself becomes a magnet, it acquires the opposite polarity. As a result it is automatically attracted. But when two identical magnets of equal strength are positioned with their *like* poles close to each other, each experiences a repulsive force equal to the attractive force that results when the two unlike poles are placed close to each other.

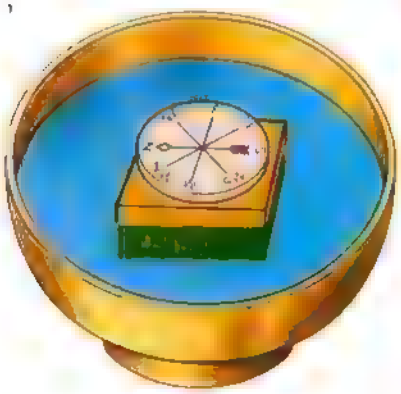
It is not only ferrous (iron-containing) materials that are affected by magnetism. But its effects are easiest to observe in pure metals such as iron, nickel and cobalt.

Domain strengths

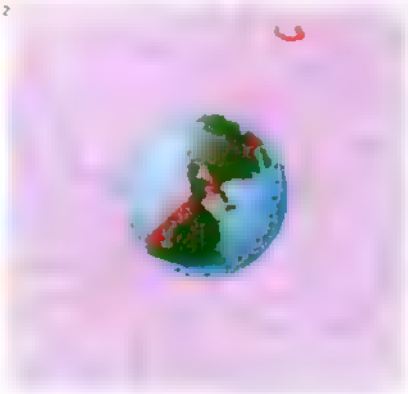
In general the metals affected by magnetism consist naturally of tiny magnets within the structure of the material, all of them aligned in a random manner. These magnets occupy regions known as domains [6], which can be seen with an electron microscope. In the unmagnetized material, the result of these

CONNECTIONS

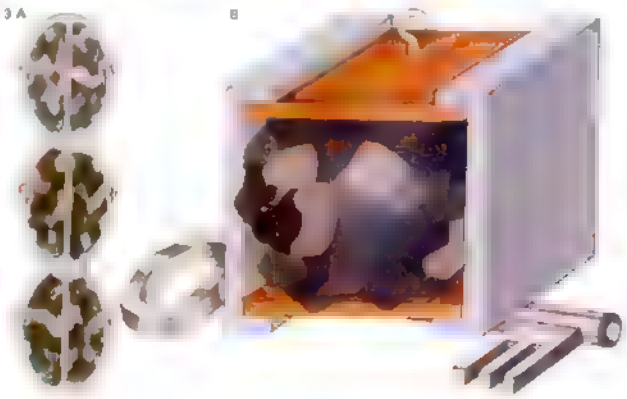
See also
Electromagnetism
Attraction and repulsion
Using magnets
Transformers, motors and dynamos
Towards absolute zero



1 The Chinese were probably the first to realize the directional properties of magnetic materials and built compasses to help them navigate at sea as well as overland. By the 12th century magnetic compasses were in use in the West. This 13th century one consists of a disc of lodestone (meaning 'way stone') marked with the compass points, mounted on a block of wood and floating in water.



2 A simplified model of the earth's magnetic field may be made by picturing a long bar magnet lying in the centre of the earth. Magnetic materials on the globe's surface tend to align themselves so that their north seeking poles point to what is called north (actually the south pole of the imaginary magnet) and their south seeking poles point to the south (north pole of our magnet).



3 In an early "magnet" the magnetic properties of lodestone were intensified by placing lumps of the material within a soft iron structure (A). This provided a path of low resistance to the magnetic flux of the lodestone and had the effect of concentrating the flux. The attractive force depends on the square of the flux density so that by directing the flux through a small section the lifting power of the lodestones was increased. Further improvements to lodestones' attractive qualities were gained using iron pole pieces (B). The flux lines follow the iron path to produce the two poles.

4 A simple way to magnetize materials such as iron and its alloys is to stroke it with a bar magnet (A), the nearness of which, coupled with its movement, tends to align the magnetic domains within the material. They then reinforce each other rather than keeping their normally random arrangements. The south seeking ends of the domain try to follow the movement of the original magnet's north pole so that the right hand side of the new magnet becomes a south pole. The domains lie with their south poles to the right, and so their north poles are to the left. Another way to magnetize a bar of suitable material is simply to hit it (B). The domains receive a mechanical shock and the earth's field tends to align them with itself. Adapting the technique shown in (A) it is possible to make a bar magnet of suitable material using two magnets (C). In this case the right hand side acquires a south pole and the left hand side a north pole.

5 William Gilbert (1544–1603), an English physician and philosopher, demonstrated magnetic phenomena to Elizabeth I. His *De Magnete* was the first major work in Europe to describe the characteristics of magnets and magnetism in an organized way. Some of the theory is now known to be incorrect but it was still the most important contribution to the subject for many years. He suggested that the earth itself is a large magnet. The compass was already at that time used as a navigational aid – an important tool for the long commercial and military ventures of the 16th century – but the process by which the compass worked had never been explained. This painting is by A. Ackland Hunt.

millions of tiny magnets, acting in different directions, is to produce a neutral field one with no magnetic properties whatsoever It is as if hundreds of children were all tugging equally at a maypole from different positions, the result of their combined efforts is that the pole does not move

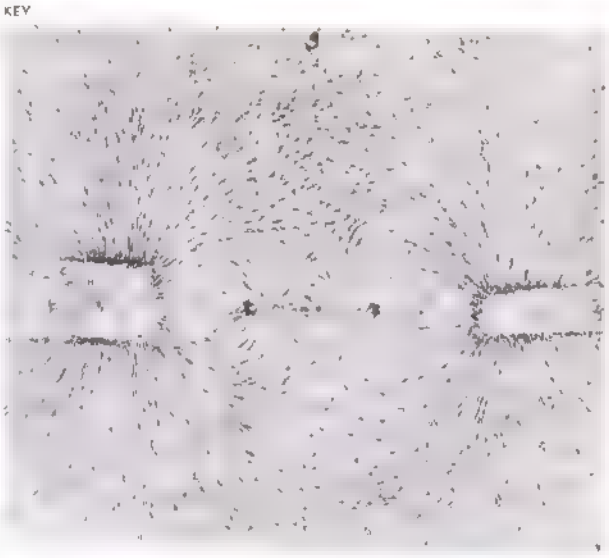
The process of magnetizing consists of causing all the domains to assist each other by lining them up in the same direction. As they all come into line the total effect is additive and the whole of the material begins to display the properties of a magnet. If all the domains become perfectly aligned then the material has reached the limit of its magnetic capability As a result, the magnetic strength of a material depends ultimately on its domain strength, and this is determined, in turn, by the way its individual atoms are structured within the domains

Earth's magnetic field

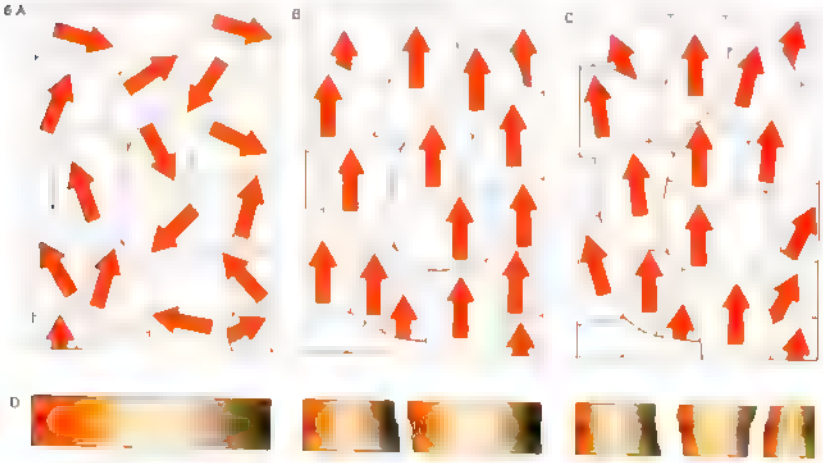
The magnetic field of the earth has been accurately measured and charted but it still cannot be adequately explained [2] In very simple terms, it is as if a single bar magnet lies between the geographic North and South Poles to produce some of the observable effects. But this does not explain the very unusual variations of strength, and even change of direction, of the magnetic forces over the earth's surface. Nor does it explain why millions of years ago the magnetic poles were oppositely aligned to their present direction, nor why they are slowly but constantly moving.

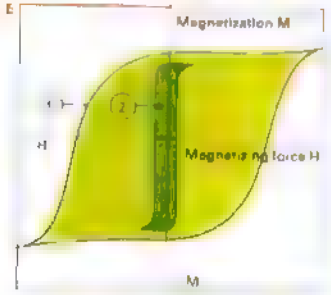
Both terrestrial magnetism and that exhibited by small pieces of iron can be better understood by considering that lines of magnetic force (often called flux lines) leave the north pole and enter the south But this is an entirely arbitrary concept, in the same way as lines of latitude and longitude on a map are drawn merely for the sake of convenience

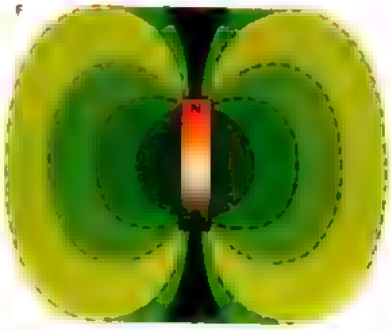
In a simple bar magnet, lines of flux [Key] are pictured as forming an approximate cylinder stretching in air from one pole to another and enclosing the magnet itself The flux lines are of the same polarity so they repel each other. They all start from and end at the same poles, but they each follow unique paths that can never cross



The pattern of iron filings in a magnetic field demonstrates how the lines of magnetic flux are distributed. The lines never cross and their mutual repulsive effects push each other away

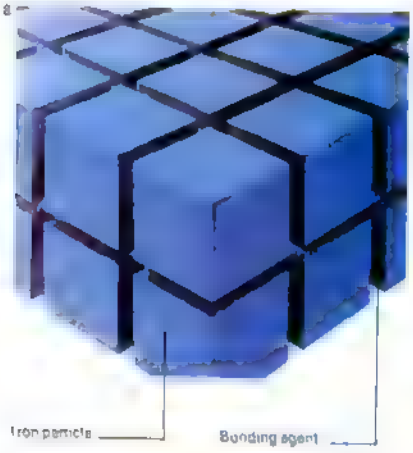






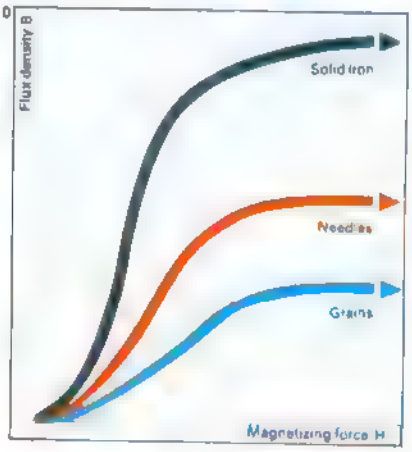
6 The random arrangement of domains in unmagnetized materials [A] becomes highly ordered in a strong external magnetic field [B] On removing the field the domains do not revert completely but retain some degree of alignment [C] Domains on each side of any breaks ensure that large bar magnets always split up into smaller replicas of themselves [D] Increasing the magnetizing force beyond a certain limit [E] cannot increase perfect alignment, and the material "saturates" Reversing H causes demagnetization Different materials have similar shaped curves [1 and 2] Removing H leaves the material partly magnetized [F]

7 Magnets can be made in almost any form, from the bar [A] to a horseshoe [B], a ring [C] or a shape like that of D, used in an electrical measuring instrument The poles are marked as N and S.



8 Magnets of complicated shapes can be made using powdered iron, mixed with a suitable bonding agent and cast into the required form. In granular form the structure is similar to bricks and mortar, the bricks being the particles, the mortar the bonding agent Each magnet is separated from the next by non-magnetic material For this reason the whole structure makes a weaker magnet than it would be if made from solid material

9 Powdered iron can also be bonded as needle like particles magnetized in such a way that their poles correspond to their points. The flux lines tend to run along the axes so that the bonding agent has a limited weakening effect



10 Solid iron is more easily magnetized than any of the forms of powdered iron (such as the grains of iron in illustration 8 or the needles of illustration 9 which are joined with a bonding agent) as the domains form an uninterrupted array

Electromagnetism

Electromagnetism is the effect by which electrical currents produce magnetic fields. Occasionally the process is unwanted, such as when a current flowing through a piece of equipment or cable on a ship produces magnetism that deflects the ship's magnetic compass. Often the effect passes unnoticed because it is very weak. But sometimes electricity is deliberately used to produce magnetic fields of great strength, as in the electromagnets that lift scrap iron [5].

Current flow and magnetic flux

The strength of a magnetic field is measured in flux lines or webers (Wb). These lines are produced whenever a current flows; and in air there is a simple proportional relationship between electric current flow and magnetic flux. A straight wire carrying a current can be looped to form a single turn. Provided the radius is reasonably small, the effect of forming a loop is to increase the concentration of magnetic flux without having to increase the current.

This concentrating effect can be further intensified by using more turns of wire to form a coil [4C]. At the point of maximum flux density – that is, maximum flux lines per unit area – the relationship between electric current *A*, turns of wire *T* and magnetic flux *B* is such that *AT* is proportional to *B*. Additional turns are simply a way of making the same current pass the same way more than once, and 12 amps flowing through three turns has precisely the same magnetic effect as three amps flowing through 12 turns.

Solenoid is the name given to a coil of wire wound to produce a magnetic field. Solenoids may be wound on iron (iron-cored) or on a non-magnetic support (air-cored). As far as flux is concerned, any non-magnetic core has the same properties as air, which means that the relationship connecting current, turns and flux holds good.

The presence of iron influences the magnetic field in two ways. It enhances the magnetic effect of the current, often by a factor of a thousand or more, but it also destroys the simple relationship applying to air-cored coils. Both these effects are a result of the structure of iron.

Microscopic regions called domains in the iron tend to align themselves with the magnetic field produced by the current. The iron provides an easy path for the magnetic flux passing through it. As a result, a given current produces more flux per unit cross-sectional area – that is, there is a high flux density. When all the domains have been aligned, further increase in current (or in the number of turns of wire in the coil) increases the flux density only negligibly.

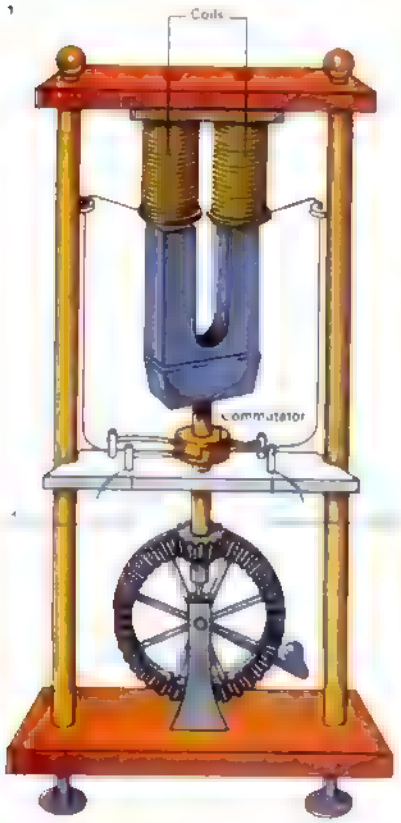
Limiting characteristics

An iron-cored solenoid has a vastly stronger magnetic field compared with that of an air-cored one but is limited by the characteristics of iron. Theoretically there is no maximum to the magnetic field produced by an air-cored solenoid. But generally the enormous currents required to make them comparable to iron-cored ones are too expensive and technically difficult to produce.

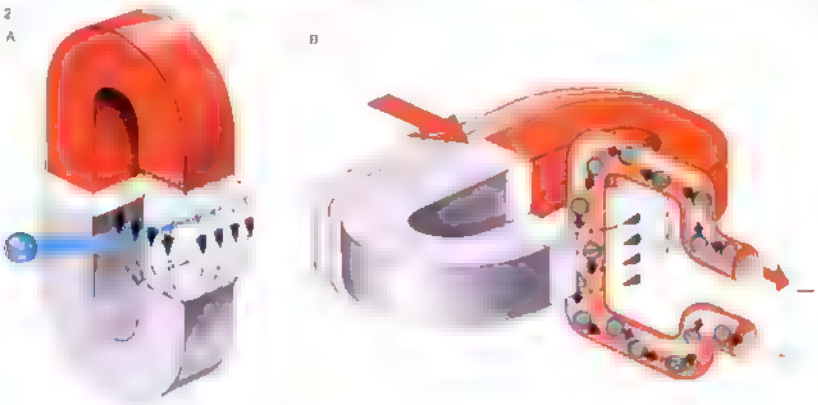
A changing magnetic field can produce a current just as a current can produce a magnetic field. As a magnet moves towards a conductor the flux lines sweeping past cause an electromotive force (voltage) to be

CONNECTIONS

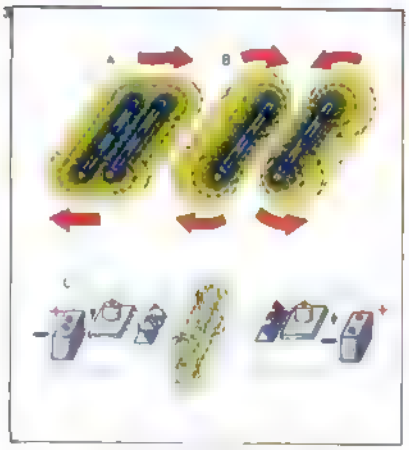
See also
Magnets and magnetism
What is an electric current?
Transformers, motors and dynamos
Static electricity
The nature of energy
Flying magnets



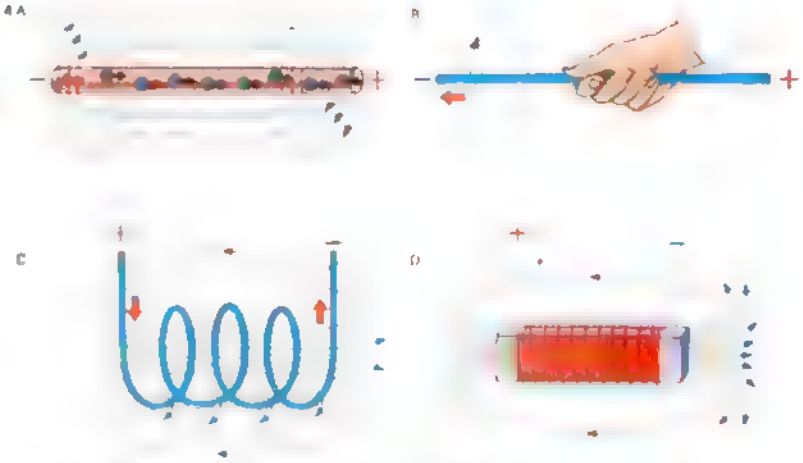
1 An electricity generator devised in 1883 by a Frenchman, Hippolyte Pixii, consisted of a horseshoe magnet set on end between two coils. The magnet was rotated through a gear system driven by a hand crank. As the magnet rotated, an alternating voltage was induced in the coils. A commutator, added later, enabled positive voltage to be picked off one side, and negative off the other, to produce direct current.



2 A moving electric charge generates its own magnetic field. This will interact with the field of a magnet to deflect the charge [A]. If the charge is moving in a wire, the wire will be deflected, as in an electric motor. Conversely, if a magnetic field is moved near a conductor, electrons will move so that their own magnetic field opposes the changing field [B]. This is the basis of the electricity generator.



3 An electric current carried by a wire sets up a magnetic field around itself. As a result when parallel wires carry currents in the same direction [A] they attract each other, but when the currents flow in opposite directions [B] they repel. The ampere balance [C] measures the force of attraction or repulsion between two electric current-bearing conductors. Such forces can be very destructive when the currents are high and the conductors close.



4 Electrons generate a magnetic field when they move. To determine the shape of the field the direction in which the "free" north pole of a small compass moves, at various points within the field, is noted. When a current flows at uniform velocity along a wire the magnetic field produced is a set of concentric rings in any plane drawn at right angles to the wire. The number of lines of force drawn – dotted lines in B, C and D – by convention, represents the strength of the magnetic field. The direction of the field can be found by applying the "right hand grip rule" [B], that is, when the wire is held in the right hand with the thumb pointing in the direction of the current flow – conventionally opposite to that of the electron flow – the curled fingers indicate the field direction. If the wire is bent the field follows the change in shape. When the wire is wound, in a coil or solenoid, the force lines of the field all point in the same direction through the centre of the coil, overall the field resembles that produced by a bar magnet with the north pole, in this example, on the right. If an iron bar is inserted into the core of the solenoid, the lines of force prefer to remain within it and this results in a high concentration of lines at the end of the bar. This device, called an electromagnet, can produce a very strong magnetic field which varies with the current in the wire, the number of turns in the solenoid and its cross-sectional area.

induced. The polarity of the induced voltage depends on the polarity and the direction of flux movement. The effect is greater in a coil than in a single wire, and is in proportion to the number of turns of wire in the coil. Similarly if the coil is iron-cored the induced voltage is more than in an air-cored coil because the flux changes are larger.

In inducing a voltage in this way there must be relative movement between the flux and the conductor (or coil). If not, flux lines will not move relative to the conductor and no voltage will be induced.

How power is produced

Electric generators produce current using precisely these principles [1]. In their basic form a magnet is rotated between coils. A voltage is induced depending on the factors outlined previously – that is the strength of the magnet and its speed of rotation (since this determines the rate of flux change). The voltage in a conductor is directly proportional to the rate at which flux sweeps past it.

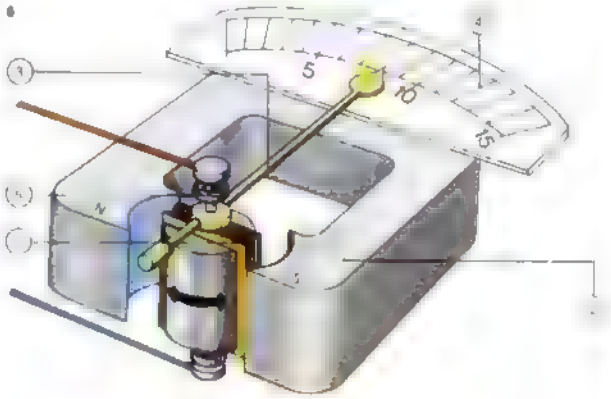
In many generators the magnet is replaced by a solenoid that must be energized or "excited" with current to produce the magnetic field necessary for the generator to function. It is the combination of voltage and current that constitutes the electrical power output from a generator.

Another aspect of the interrelationship between current in a conductor and magnetic flux makes use of the flow of an electric current in a magnetic field to produce physical movement. This is the principle on which motors and some electrical measuring instruments operate [6] but electrical power must be supplied to cause movement against a mechanical force.

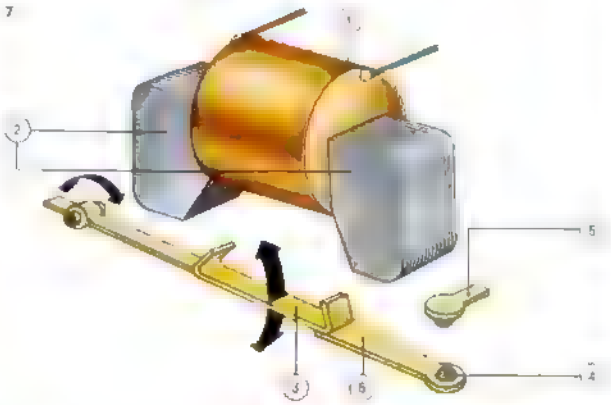
Magnetic fields far stronger than ever before are now created by means of superconductivity, the zero-resistance effect in some metals at temperatures approaching absolute zero. As a result, current can flow without losses or heating, and it is possible to use vast currents in air-cored coils, avoiding the limitations of saturation imposed by iron. These enormously strong magnetic fields open up prospects for electromagnetic levitation and new forms of motors and generators capable of high outputs at reduced costs.

The 'lift' obtainable using magnetism has been applied to make a working model of a magnetic levitation train which is seen here end-on. The train has no wheels, but instead "floats" over a long magnetic strip that takes the place of a conventional track, below which a series of electromagnets generate the necessary magnetic field. The electric current can be supplied to these electromagnets in such a way that they behave as a linear motor, so driving the train along the track. Such trains are frictionless, pollution-free and virtually silent. Experimental prototype trains have been tested in Britain, West Germany and Japan.

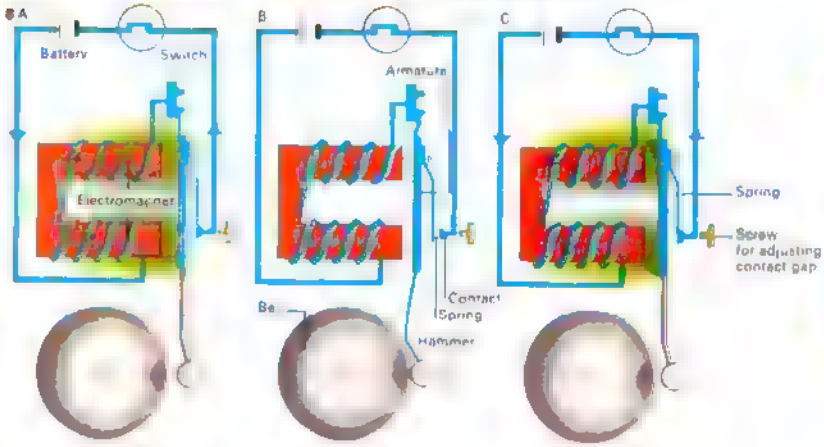
5 Electromagnets are often used in scrap metal yards to lift ferrous metals. This method not only reduces manhandling but also provides a means of separating the iron from other scrap materials.



6 In electrical indicating instruments a coil [1], turns when energized in a magnet's field [2]. The pointer [3] shows the strength of current on the scale [4]. The hairspring [5] returns the pointer to rest.



7 Relays are used to allow a low-power source to switch a high voltage circuit. When the coil [1] is energized by a small current, magnetic flux appears between the poles of the core [2] attracting the armature [3]. The moving contact [4] then engages the fixed contact [5], closing the high-voltage circuit. When the coil is de-energized the balance weight [6] overcomes the weakened magnetic field and opens the circuit.



8 In an electric bell, the magnetic field of an electromagnet is effectively turned on and off rapidly to make a hammer strike the bell. Pressing the switch [A] allows current from a battery to energize the magnet, pulling over a spring loaded armature and with it the hammer [B]. This action also breaks the circuit at a point contact and "turns off" the electromagnet. The armature springs back, re-makes the circuit [C] and the whole sequence is repeated.

9 A loudspeaker commonly has a permanent magnet [1] to provide a magnetic field in which a coil [2] attached to a fibre cone [3] is held balanced but is free to move backwards and forwards. A varying electrical current is fed to the coil from the amplifier output, resulting in a varying magnetic flux in the coil. This reacts with the permanent magnet's field, causing the coil and hence the cone to move back and forth, producing sound.

Using magnets

The two main types of magnets used in apparatus such as electric bells, motors, dynamos, speedometers and the like are permanent magnets and electromagnets. Permanent magnets, chiefly made from iron-based alloys, retain their magnetism all the time. An electromagnet consists of a coil of wire, sometimes wound round a soft iron core, and behaves as a magnet only when an electric current flows in the coil.

The magnetic field

Those processes that generate motion by creating a strong and then a weak magnetic field (as in an electric bell) expend more electrical energy in the electromagnet than is gained from the resulting mechanical motion. But weakening a magnetic field (or alternatively "turning it off") is in fact the basis of many important devices that have characteristics that would be difficult or expensive to obtain in other ways. In permanent magnets the field cannot be turned off without destroying it. But it can be diverted.

The best example of diversion of the magnetic field is the magnetic chuck [Key]. This is a device for holding ferrous metals tightly onto a work table. The chuck is used almost exclusively in grinding machines because a vice could distort the metal, or not hold it level relative to the grinder.

The chuck comprises a number of small bar magnets embedded on a movable metal plate so that north and south poles are pointing vertically. The metal plate is of a material with low magnetic properties and the poles of the magnets are positioned alternately north–south and slightly separated. A second metal plate is placed above the magnet assembly. This incorporates soft iron pieces that correspond to the position of the magnets fixed in the base plate.

When a workpiece is placed on the upper plate it provides a flux circuit for the embedded magnets through the soft iron pieces of metal and is attracted to them, holding the workpiece in position. However, the operation of a simple lever allows the lower plate to be moved horizontally in such a way as to bring the magnet poles out of alignment with the pieces of soft iron in the upper plate. This causes the magnetic flux to be diverted from the workpiece and it goes instead through the metal of the upper plate to link the embedded magnets north to south. The workpiece is then free to move.

Magnets in railway service

A system in which permanent and electromagnets complement each other for reasons of safety is commonly used in railways. A strong permanent magnet is attached close to the track at a set distance from the signals. As a train passes over the magnet it causes a pivoted permanent bar magnet in the cab to swing (like a see-saw) through a small angle and rest in this new position. The bar magnet's movement closes a switch to bring into action an alarm bell or hooter signal. A few seconds later the cab passes over an electromagnet connected to the signals. If they are set at "clear" the electromagnet is energized and the pivoted magnet in the cab is repelled so that it returns to its first position, turning off the alarm.

But if the signals are set at "stop" or "caution", the electromagnet is not energized and, after a short pre-set delay, the brakes

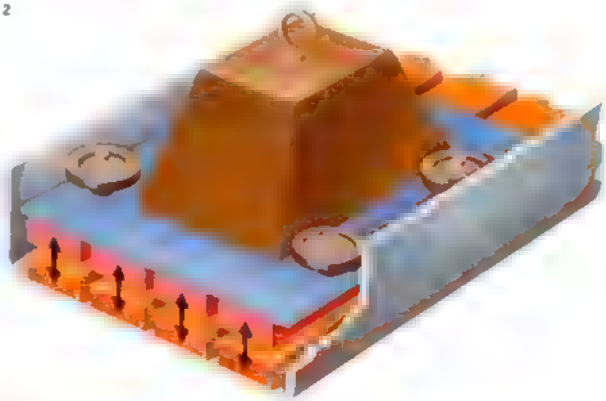
CONNECTIONS

See also
[illegible]

1

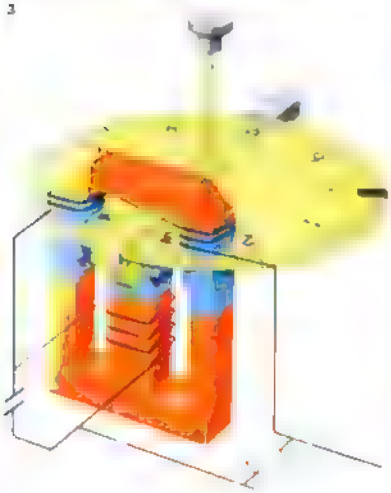
1 Magnetic mines took a great toll of merchant ships during World War II. Placed in busy shipping lanes, the strong magnets built into the devices were attracted to metal hulled ships which caused the mine 4 to explode. Counter measures such as electrical cables to reduce the ships' magnetic fields were devised and are sometimes used in places where mines may still be located.

2

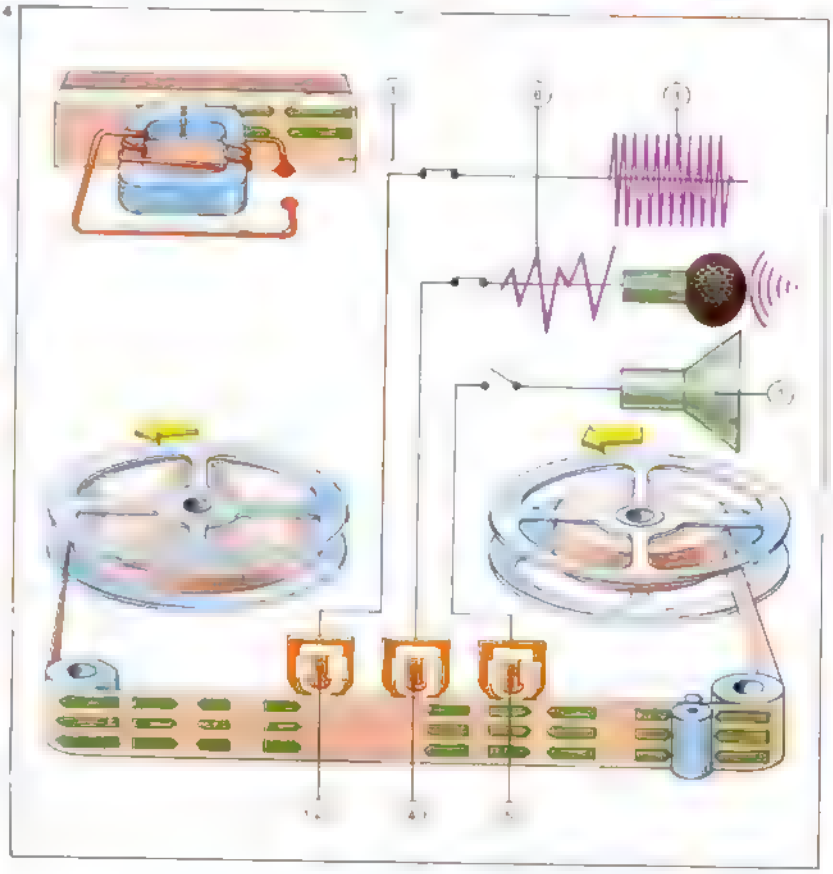


2 A system of permanent magnets attached to movable metal plates can be suspended above a fixed set of magnets. Provided there is a guide rail to stop sideways motion, it can be used to move heavy loads around a factory or within an area where it is convenient to build a magnetic "track". The advantage over comparable methods such as rail systems is the absence of friction and moving parts.

3



3 A domestic electric meter – a watt hour meter for recording the amount of electricity consumed – makes use of electromagnets. When current is being used, it flows in coils that energize magnets and make a disc rotate. The disc is coupled to a counter reading in kilowatt-hours.



4 Magnetic metal oxide development has led to a revolution in the sound-recording industry. Metal oxides in powder form are bonded to flexible plastic tape [1], usually PVC, forming a moving surface on which magnetic patterns can be imposed corresponding to sound, visual or other signals. These tapes are used in machines that consist of an erase head [2] using high-frequency input [3] to demagnetize the tape as it is driven past, and a record [4] and replay head [5]. These either magnetize the tape according to the signal input [6] or reconvert the previously imposed magnetic patterns into the signals that formed them (ie plays back) [7]. Stereo or twin-track recording combines two record/replay heads in one. The erase head is taken out of circuit when playing back and, to prevent accidental erasure, most tape recorders have a built-in fail-safe arrangement.

are automatically applied if the driver fails to apply them. The brake-time circuit (like the audible alarm) is energized from the moment the pivoted magnet moves. If this magnet is returned to its original position (within the pre-set time) the brakes are not applied.

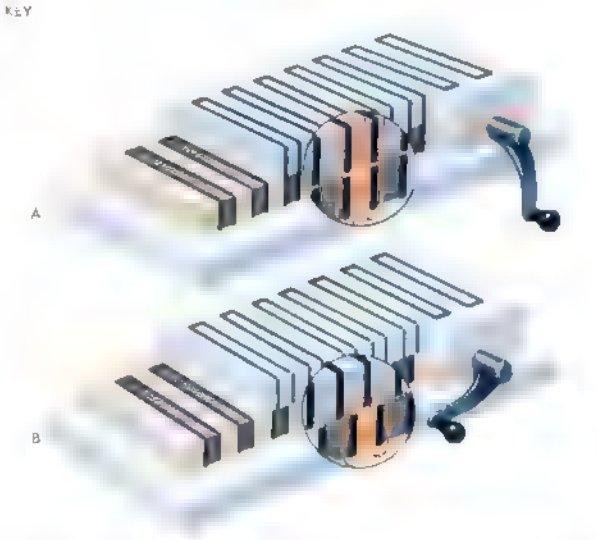
Meters and medicine

A phenomenon associated with magnetic fields is the eddy current. When there is relative motion between an electrical conductor (not necessarily one with magnetic properties) and a magnetic field, currents called eddy currents are induced. These, in turn, produce another magnetic field of opposite polarity. There is a tendency, because of the attraction between the opposite fields, for the conductor and original magnetic field to move together while the motion exists.

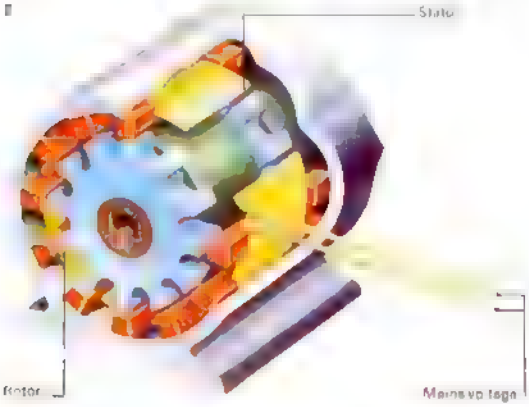
This principle is the basis of the car speedometer. A permanent magnet within the instrument housing rotates at a speed related to the crankshaft speed. It turns within a specially shaped aluminium disc which itself can rotate, but only through about 270° because of the restraint of a spring. As the magnet rotates, eddy currents are induced in the disc, which tries to follow the magnet. The strength of these currents is proportional to the speed of rotation and therefore the disc moves according to the speed of the car. A pointer attached to the disc moves over a scale calibrated in kilometres or miles per hour.

An electricity meter [3] works on similar principles. Current used by the consumer passes through an electromagnet, which induces eddy currents in an aluminium disc. In this case the disc can rotate freely through 360° and its movement is coupled to a gear train that drives the indicating dials.

Medical science also benefits from the use of powerful magnets. Experiments are taking place in the application of magnetically guided "pills" within the body. The "pill" which may be swallowed or inserted in a vein is a minute radio, capable of transmitting information about such factors as temperature and salinity. It incorporates a "nose" of suitable metal and may be guided to particular organs by a magnet operating outside the body, thus aiding doctors in diagnosis.

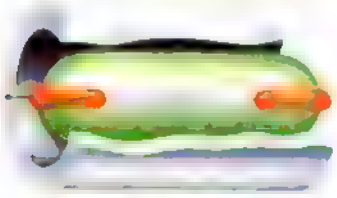


A permanent magnet cannot be switched off as electric currents can. Instead its magnetic flux can be diverted. A magnetic chuck holds a workpiece in position on a grinding machine. Flux enters the workpiece and holds it in place [A]. When the flux is cut, the work is released [B].

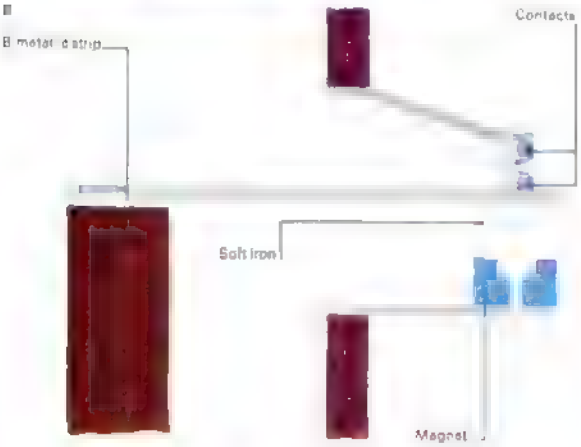


5 Magnetically polarized rotators are essential elements of modern timing devices and similar equipment. Because the mains frequency is fixed precisely at the power station, motors whose rotation is tied to this frequency have a highly accurate speed. Using magnets in the rotor (often permanent for ease of manufacture and reliability) ensures that once it is "locked" to the stator frequency, the rotor follows the speed exactly. The rotor is made from thin sheets of silicon steel cemented together to form a cylinder. The ridges are then magnetized to form poles.

6 Electricity meters are often difficult to read. The cyclometer type shown here suffers from the disadvantage that the digits representing the high numbers (on the left) rotate slowly, to give ambiguous readings. A version developed by Ferranti has a small magnet attached to every wheel behind the figures 7. A bar magnet is fixed over the wheels as shown. When the wheels move from 9 to 0 magnetic attraction ensures a quick changeover of figures. Only the highest number wheel is excluded from this arrangement because when it changes to 0 all the other figures will change to 0.



7 The reed switch, used as a safety switch and in electronic counters, has its contacts enclosed in a glass envelope to protect them from corrosion. A magnet is used to close or open them.



8 A thermostat often incorporates a bimetallic strip, which moves to open or close a circuit according to the temperature. One way of ensuring a quick "on" or "off" is to use a fixed magnet and a soft iron pellet (armature) attached to the moving contact arm. As the bimetal starts to curve it is either attracted swiftly to the magnet or else is released suddenly, providing a fast "snap" action.

9 The electromagnetic clutch, often used on ships, has a soft iron cup [1] attached to the propeller shaft and has electromagnets on the engine shaft. Energizing the coils produces a strong field and transmits the drive motion.

10 Magnetic catches are used on doors of furniture and refrigerators to hold them shut. A common type on a refrigerator uses "magnetic rubber" (rubber with ferrous particles magnetized to form a convenient pole pattern) on the edge of the door. The type shown here, on a cupboard, uses metallic magnets in a plastic housing fixed to the frame. It "catches" on an iron plate on the door.

Transformers, motors and dynamos

The transformer, one of the most essential and efficient of electrical machines, has wide uses in the supply of electricity. It is used in power stations and at sub-stations – in the former to boost voltages for transmission over power lines and in the latter to reduce voltages to levels suitable for industrial or domestic use. Transformers are also used in many electrical appliances – such as radios, television sets and battery chargers – wherever alternating voltages different from the mains supply are required.

Motors and generators

A transformer, with its two main elements, magnetic and electrical, linked by a laminated soft iron core [Key, 1], has no moving parts and is up to 99 per cent efficient. Electric motors and dynamos, however, have rotating elements and therefore cannot be made quite as efficient.

Motors and dynamos (or generators, as they are now more usually called) are basically the same in construction, although their functions are different. Motors are supplied with electrical power to provide mechanical power; generators are supplied with mechanical power to give electrical power. But it is important to remember that they are so similar that some machines can act as motors or as generators, depending only upon whether they are supplied with electricity or mechanical power.

The two essential elements of each of these machines are the field and the armature. The field is a magnetic field, which may be derived from permanent magnets or electromagnets. The former are cheaper but the latter are more convenient because with an electrically energized field it is easy to increase or decrease their strength. It is, however, a convenience gained only at the expense of having to provide the coils for the electromagnets that form the field (field "windings") and to do all the work that goes into insulating and installing them.

How the armature works

The armature is also a winding but is arranged differently from the field. It is essentially a conductor (or conductors) arranged to cut the field's magnetic lines of flux at right-angles. The armature conductors may be wound onto a cylinder that rotates in a field. Or they may be fixed to the inner walls of a cylinder, within which the field windings rotate. The static part of the machine is called the stator and the revolving part the rotor. Both the field and the armature may be on either the stator or the rotor.

By a basic principle of electromagnetism, a voltage is induced in a conductor that moves in a magnetic field and a conductor in a magnetic field experiences a force and tends to move when a current flows in it.

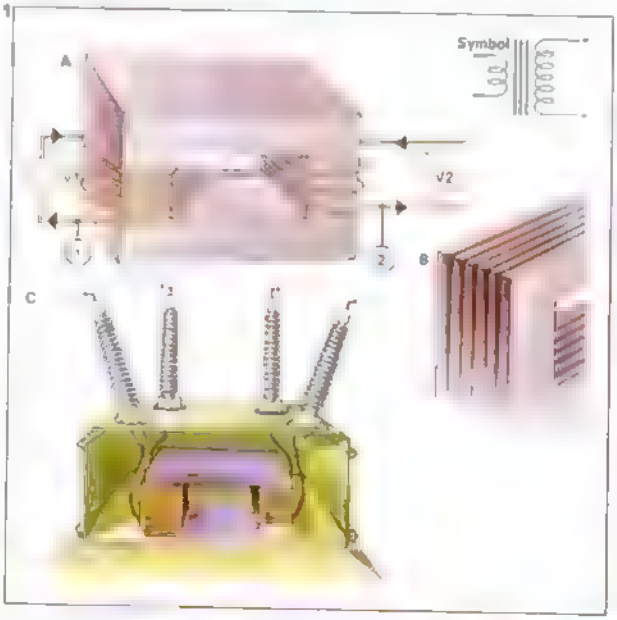
To make the best use of this basic effect, the magnetic and electrical elements in electrical machines (the field and the armature) have to interact in the most efficient manner possible. The armature is generally wound on an iron core which concentrates the maximum number of magnetic flux lines. The field coils are also wound on iron cores, to produce the maximum flux for a given current. The iron for both the field and the armature is sometimes laminated – that is, made up of slices [1]. This prevents the currents from circulating and "eddying" in the iron.

CONNECTIONS

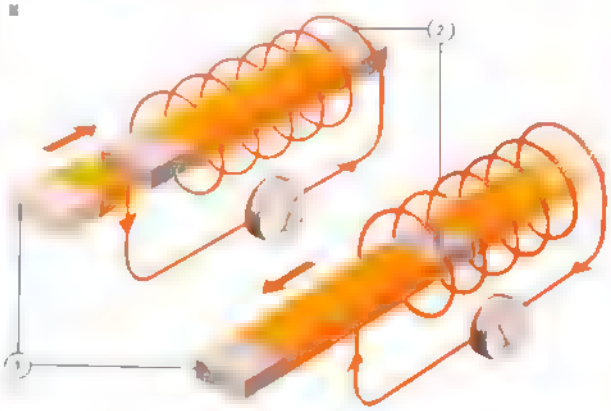
See also

Magnets and

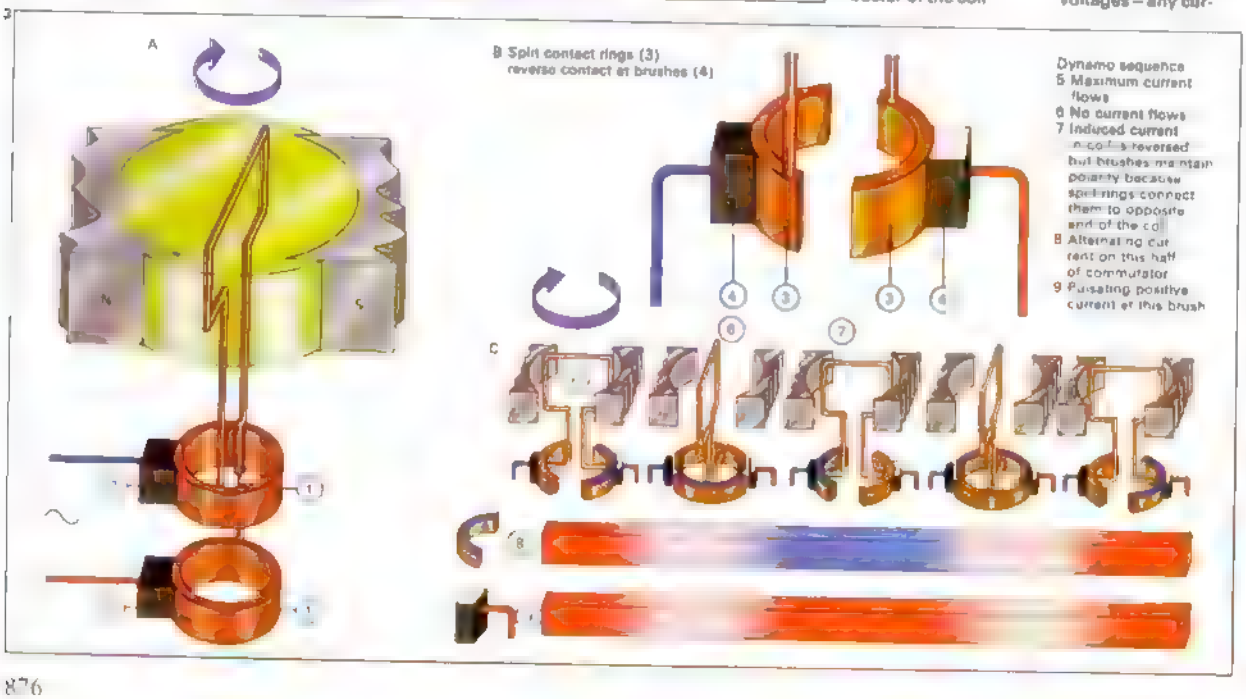
What is an electric



1 In the transformer [A], the input or primary current [1] causes lines of magnetic flux to form in the iron core, linking it to the output or secondary [2]. As the supply alternates the flux lines collapse and reform in the same pattern, but with different polarities. They induce a voltage in the output coil. The ratio of input to output voltage (V1 to V2) is the ratio of turns on the input and output coils. The iron core is laminated [B] to reduce eddy currents. The high-voltage transformer [C] has its terminals insulated to prevent flashover.



2 Electricity can be generated by a magnet, an electrical conductor and relative movement. Moving the magnet [1] causes the flux lines surrounding it to cut the conductor of the coil [2] and induce a voltage in it – coinciding with the movement: the faster the movement the higher the voltage induced. Opposite movements produce opposite voltages – any current in a circuit between the coil ends flows first one way then the other as the magnet is moved in and out. These are alternating currents and the generators are called alternators.



3 Electricity is generated mechanically when a coil is rotated in a magnetic field [A]. An AC voltage is induced in the coil and is connected to the external circuit by contact rings [1] and carbon brushes [2]. A current flows when a circuit is made between the brushes. To produce direct current – as in a cycle dynamo – the generator is modified [B], the contact rings are split [3] and the halves insulated from each other. A fixed pair of brushes [4] contacts the ring segments in turn. The arrangement of split segments is called a commutator: the sequence [5–9] describes how it works in a cycle dynamo.

itself and generating wasted heat, as a result of changing magnetic flux in the machine.

The armature must be supplied with current (through rotating contacts) if it is the rotor in a motor and there must be a way of taking current from it if it is in a generator. The same applies to the field if it is electrically energized. The rotating contact arrangement can be either a set of slip-rings or a commutator [3]. These rotate under fixed contacts, called brushes, made of carbon and held in place by springs. From time to time the brushes have to be replaced when the carbon wears away.

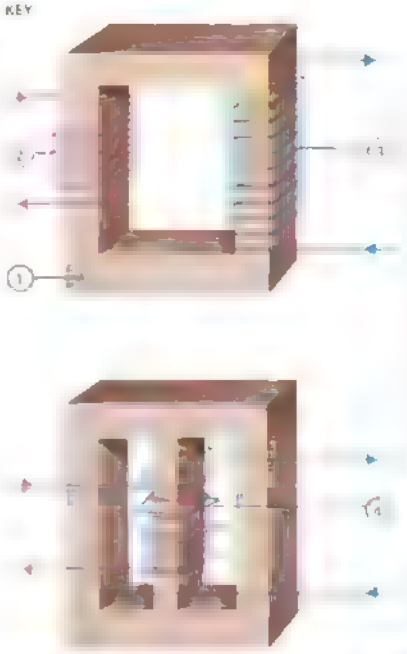
Varying the field

The armature and field windings of a motor are supplied with current and, by using a resistor in the field circuit, it is possible to vary the field strength. Weakening it causes the motor to revolve faster – provided the armature current is constant – but with less torque (turning force), and vice versa [4]. In a generator driven at constant speed, strengthening the field increases the voltage output; weakening the field decreases it.

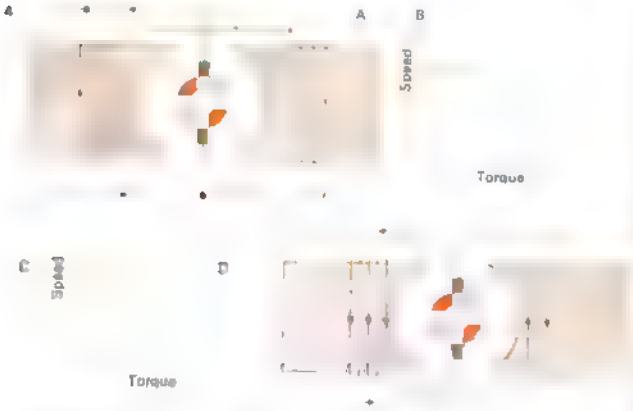
The motors and generators already described are generally suitable either for AC (alternating current) or for DC (direct current) supplies. However, induction motors can be used only with alternating current supplies. The armature is the rotor and can be wound or made in the form of a "squirrel cage" [5]. The flux from the stator field is changing and induces a voltage in the cage. This causes a current to flow which interacts with the stator field to produce torque. No slip-rings are necessary.

The simplest field consists of two coils. As one becomes a north magnetic pole, the other becomes south, unlike DC machines whose field polarities are constant. The rotor experiences a pulsating and reversing field but not one that appears to rotate, without a starting force the armature will not rotate either. Given a push in either direction, it begins to follow the field as it alternates.

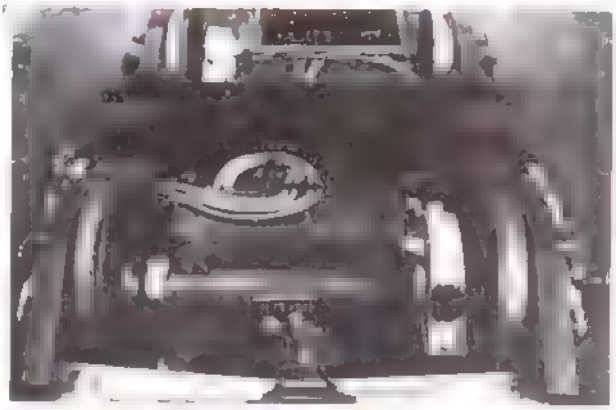
Some induction motors have a wound rotor and slip-rings may be used. The purpose of the rings is not to supply current to the armature but to alter its characteristics with the use of external resistances.



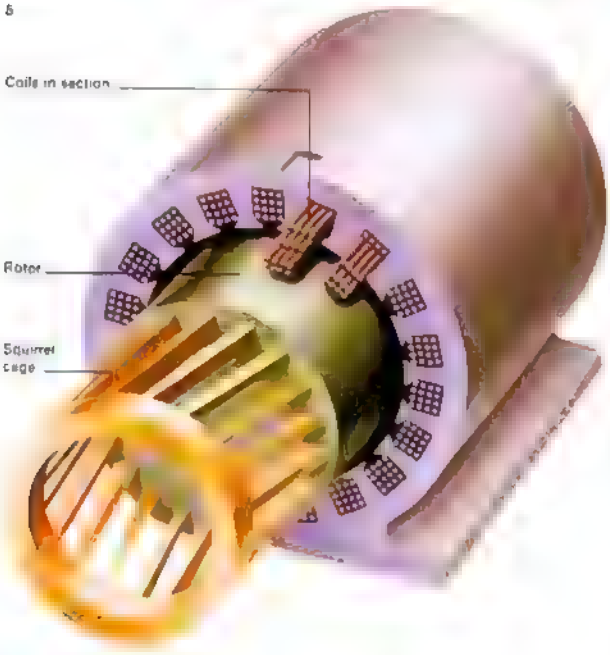
The transformer a simple and efficient means of raising or lowering AC voltage, consists of three basic elements: an iron core [1], which provides a magnetic link between the primary or input [2] coil and the secondary, output coil [3]. The turns ratio between the input and the output determines the ratio between AC voltages: 'transformed'. Fewer turns on the output side give a proportional decrease in voltage and vice versa. Transforming the flow of current. The core may take a number of shapes, sometimes more than two coils are used and sometimes the two coils are wound on each other [4].

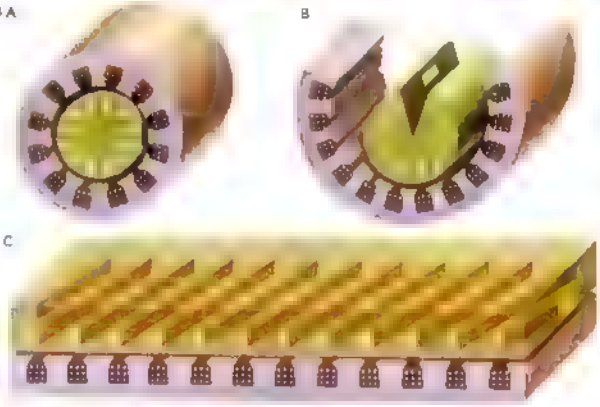


4 Electric motors are similar in principle to generators. Current supplied to the armature or coil and to the electromagnetic field, causes the armature to rotate. Connecting the field coils and armature in parallel [A] gives an almost constant speed for any torque [B]; connecting them in series [D] produces high torque [C] at low speeds, as in motors [E] for starting electric trains.



5 The induction motor is the one in most widespread use, the coil forming the armature of a simple DC machine is replaced by a "squirrel cage". This consists of aluminium or copper bars connecting each end to a ring, the whole embedded in a laminated iron rotor. The field at least two coils inside the motor body is shaped to allow the rotor to revolve inside with a small clearance. Flux lines caused by an alternating current passing through the field cut the cage bars inducing a current in them – hence the "induction" motor.





6 The "rotor" of a linear motor moves length-wise rather than revolving. It is a flat plate either sandwiched between two long field windings or resting solely on one. Energizing the fields with alternating current causes the plate to move in a linear direction, making use of exactly the same principles as an induction motor [A], from which it can be pictured as being "made" by cutting [B] and opening out [C].

7 A large linear motor, with field windings similar to those shown here, could be used to drive a silent, non-wheeled train or other vehicle. Smaller motors are already used – for example for opening and closing sliding doors. The armature or metal plate, usually aluminium, is fixed to the upper part of the door and the windings attached to the door frame. When the field is energized, the plate passes horizontally along the field, thus moving the door.

Basic DC circuits

An electrical circuit is the system by which an electric current is directed, controlled, modified, switched on or switched off. Circuits can contain from two or three to many hundred different components, according to the way in which the current is to be controlled but all share certain characteristics

The formation of a circuit

The primary requirement of a circuit is that it must form a complete path, electrons must be able to flow round the whole system so that as many electrons pass back into the source of the current as leave it. Certain occurrences, such as lightning strikes or electric shocks, seem to deny this first requirement, but are nevertheless examples of electrical circuits This apparent contradiction can be resolved by considering the earth and all the structures on it as a vast electron bank If clouds develop an electron imbalance, the earth makes it up with a flash of lightning and the net result is that on average the numbers of electrons leaving the earth and arriving at it are equal

Electric currents can also be "carried" by charged atoms, or ions. Ions of dissolved salts and other chemicals conduct current through the electrolyte in an electroplating bath and gas ions conduct electricity in a fluorescent strip-light. But whatever the current carriers, all circuits share three characteristics: a current (I), a voltage (V) and a resistance (R)

An illustration of electron movement (and the use of high-voltage direct current) is the arrangement of the transmission line from the Cabora Bassa dam in Mozambique to the South African town of Apollo some 500km (310 miles) away. There are two lines to carry the current, one taking electrons to Apollo, the other returning the same number to Cabora Bassa If one of the lines breaks, the earth itself "replaces" it and carries the electrons in the appropriate direction

In a similar way, the chassis of vehicles are used as the so-called earth or return circuit, although this is a loose and generally inaccurate term. One terminal of the battery is connected to the bodywork and a single wire is brought from the other terminal through a switch to each piece of electrical equipment These in turn are also connected firmly to the chassis The circuit so established allows the number of electrons leaving the battery to be matched exactly by the number of those arriving. Using the chassis to complete the circuit in this way makes a second wire to each component unnecessary.

Direct and alternating current

One major practical difference in circuit components (although there are no differences in principle) is determined by whether they are used for direct current or alternating current. Direct current is unidirectional, the electron flow is always in the same direction and although it may stop and start, grow or diminish in quantity, it never reverses

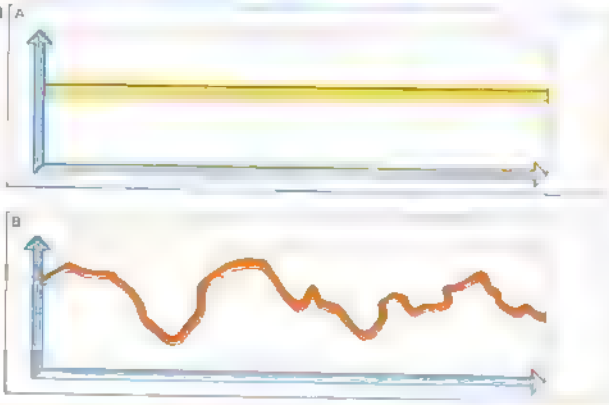
Current flow (as opposed to electron flow, which is always against the conventional direction of current flow) is assumed by convention to be from a positive to a negative terminal In direct current (DC) generators, batteries and some other sources the terminals are determined by the nature of the machine or equipment and are irreversible. The most common example of direct current source is a chemical cell or battery, in which the nature of the chemicals themselves

CONNECTIONS

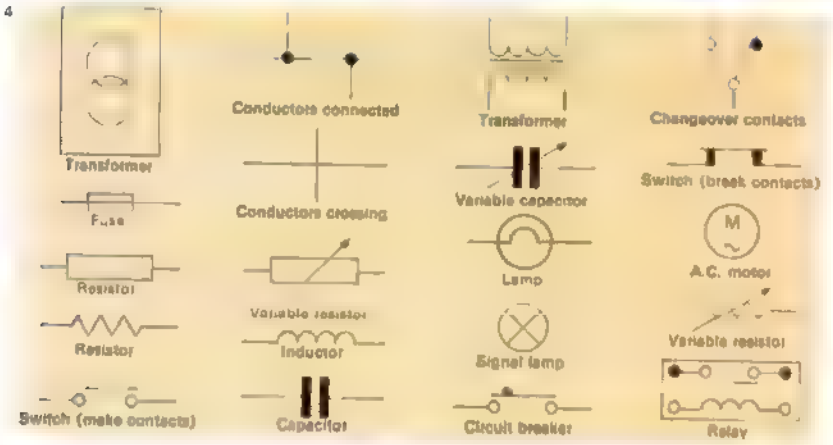
See also

1 A cell or battery sends electrons round an electrical circuit as a result of chemical action. This unidirectional character of the electron flow is called direct current (DC) even if it varies greatly or stops altogether from time to time In a primary cell current flow stops as soon as one of the chemicals or electrodes is consumed A secondary cell or accumulator can be recharged, often forming these gas bubbles

2 Electrical generators can be AC or DC machines. The output of a DC machine does not depend on an exhaustible chemical process, so it usually gives a steadier direct current than chemical sources. Large generators like these supply alternating current (AC), characterized by a rapid periodic reversal of electron flow In an AC system the current falls to zero every time the direction of its flow is reversed



3 A method of examining direct current is to measure how it varies over a fixed period. Graph A shows a current, such as that taken by a lamp supplied by a generator, which does not alter over the time it is measured (the vertical axis represents current and the horizontal one time). Graph B shows a direct current typical of a welding circuit It varies with time, although its value is always positive.



4 The drawings used by engineers and other workers in the electrical industry have to be clear to the reader irrespective of his language These symbols are just a few of the hundreds used

5 It is often economical to use direct current at high voltage when transmitting large amounts of electrical power Only in the last few years has reliable equipment been available to switch large DC voltages on and off This large thyristor "switch" operates on similar principles to transistors used in radios Because power stations are being sited farther away from population centres high voltage DC is increasingly used instead of AC

fixes the polarity of the system. Although the output of the cell may vary, the current flow is always in the same direction. The same applies to a DC generator (dynamo), because the structure of the machine determines the polarity. Another device, known as a rectifier, also has fixed polarity. A rectifier is used to convert alternating current to direct current – irrespective of how the input varies, the current direction at the output terminals is always the same.

Alternating current (AC) is the more common mode, although in certain instances direct current is particularly appropriate and alternating current cannot be used. In electroplating, for example, direct current is used because it is vital that the current always flows in the same direction. If it did not, material would pass back and forth from the coating metal to the coated surface and no plating would take place. The recharging of a battery [1], which is a specialized form of electroplating, can also be carried out only by direct current and mains-powered battery chargers must contain a rectifier.

In systems for transmitting power over long distances, direct current may be chosen because it requires less insulation and uses fewer and narrower conductors for the amount of power transmitted than does alternating current. Direct current, provided it is steady, uses its conductors fully throughout the whole of the transmission period whereas the conductors of an alternating current are not always fully utilized. As a result, a smaller conductor can be used for the same effective power transmission and if the current is to be transmitted over very long distances the savings may be considerable.

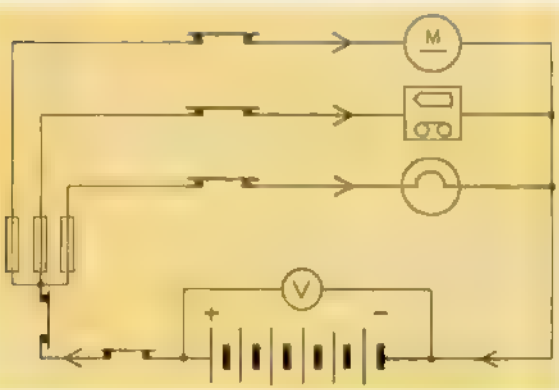
Interrupting the flow

Despite some advantages, the use of direct current is beset by one major problem: switching it off quickly. Interrupting a current that is flowing tends to produce a spark between the contacts providing the interruption. Sometimes this can take the form of a spark so large that it melts not only the contacts but the switching device itself. Alternating current, by its very nature, falls to zero many times a second, so that the spark also falls to zero, limiting the damage it can do.

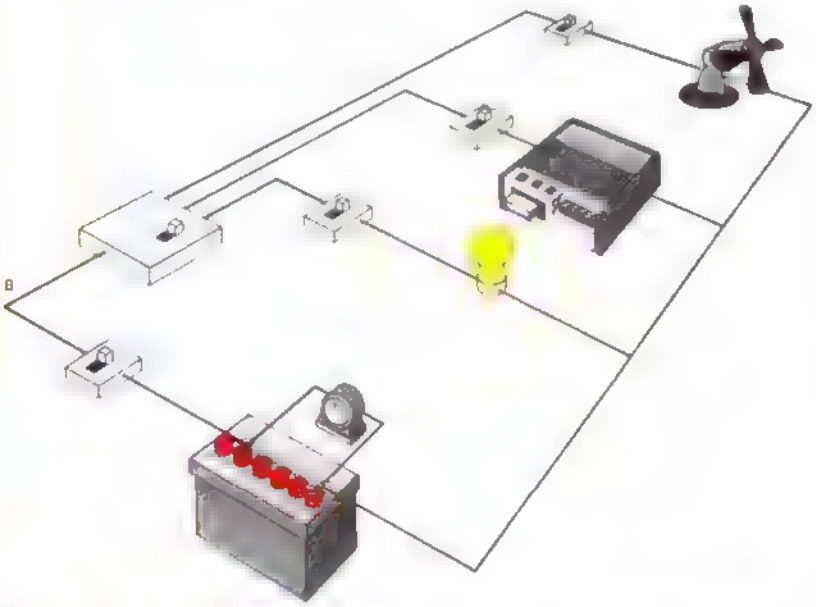
KEY

Electroplating of metals was one of the earliest uses of direct current. The inherent electrical properties of certain chemicals in solution make it possible to coat metallic surfaces with a thin durable coating of another metal. In this automatic plating plant car components are given a protective coating of chromium.

6 A

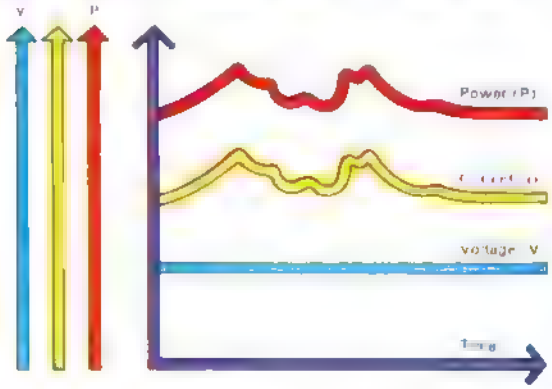


B



6 This diagram of a circuit [A] shows a typical but simplified electrical system [B], of the kind that is used for equipment powered by a battery. Switches control the current flow into three leads to a lamp and two appliances containing electric motors. The voltmeter records the electromotive force (that is, the voltage) of the battery, which remains virtually constant. The current flowing through the lamp will also remain virtually constant but those through the appliances may vary, depending on the demands made by the motors in them. Fuses can protect the circuit from current surges.

7



7 The circuit shown in illustration 8 has many interdependent electrical variables. Representing them in the form of a graph indicates how they are related. There are different scales on the vertical axes but the horizontal axis, time, is the same for all the other variables. Three of the most important variables – voltage, current and power – are shown. The current (I) is fixed by the battery. If the battery is in good condition the voltage will not vary significantly. The voltage (V) depends on the voltage and the resistance of the appliance it is feeding. This may vary (as it does in the motors) or may be fixed (as it is in the lamp). The power (P) is the product of voltage and current and is measured in watts.

8

8 Welding is a process of joining metals together with a bond so strong that it is often superior to the parent materials. A large electric current, perhaps 2 000 amps at low voltage, can be used to provide the necessary heat, so that the metal in a particular area is melted. The current passes in the form of an arc from an insulated electrode in the welder's hand and enters the work piece across a short air gap. The object that is to be welded is connected to one terminal of the current source and the welding electrode to the other. Electric welding generally uses direct current generated by special equipment because of the very large currents needed.

Basic AC circuits

The physical processes that take place in electrical circuits carrying alternating current (AC) differ from those in direct current (DC) circuits, reflecting the differences between the two types of electricity. Alternating current regularly reverses its direction, becoming zero before each reversal (100 times a second in European countries and 120 times a second in North America, corresponding to 50Hz and 60Hz supply frequencies). With reference to zero, the current is negative and positive alternately. Direct current always flows in the same direction

Current wave and circuit components

The shape of the "current wave" (the curve representing its change of value with respect to time) can take an infinite variety of forms. For most purposes it is sinusoidal (like a sine wave [Key])

The number of times the curve repeats the whole alternating cycle in a second is called the frequency and is measured in hertz (Hz) – one cycle per second equals 1Hz. A sinusoidal voltage (V) applied to a circuit produces a sinusoidal current whose value at any instant in the cycle is equal to V/Z, where Z is called the impedance (which depends on the resistance, capacitance and inductance of the circuit and the supply frequency); Z is measured in ohms (Ω). The equation is analogous to that used to express Ohm's law the direct current flowing through a conductor is directly proportional to the electromotive force (voltage) that produces it and inversely proportional to the resistance

The three main types of circuit components are inductors, capacitors and resistors. A resistor behaves in the same way in either an AC or a DC circuit; inductors and capacitors, however, do not. In these devices, currents are out of phase with the applied voltage in parallel circuits (in which there is more than one path for the current) and voltages are out of phase with the current in series circuits (in which the source and output devices are connected by only one path)

Phase lead and lag

A simple analogy to the phase differences in alternating current and voltage is the action of a yo-yo, where the hand from which the spinning mass derives its energy can move in the opposite direction to that of the mass The current taken by a capacitor is out of phase with the applied voltage; it is zero when the voltage is maximum and vice versa Sine waves may be represented by rotating vectors (a vector is a quantity that has both magnitude and direction) and on a vector diagram the capacitor current is 90° out of phase with the voltage and is said to be leading.

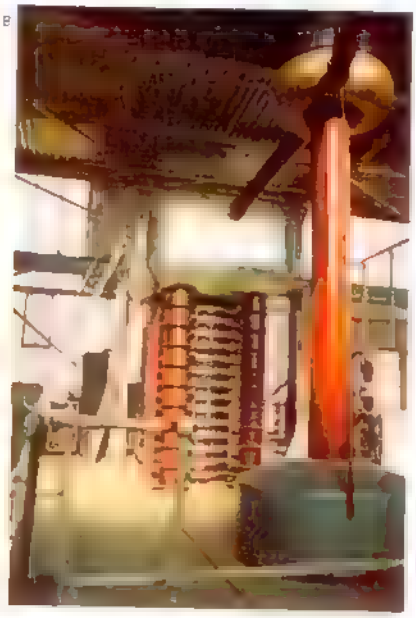
For a pure inductance the reverse applies – that is, the current lags the voltage by 90° This can be explained in another way by saying that for a capacitor the voltage lags the current by 90° and for an inductor the voltage leads. With a resistance, the current and voltage are in phase [3]

In a circuit that has both capacitance and inductance but no resistance, one current leads by 90° and is equal in magnitude to another lagging by 90°. The overall effect is subtractive – they cancel each other out When this happens the circuit is said to experience current resonance. In effect, the capacitor's current feeds the inductor and

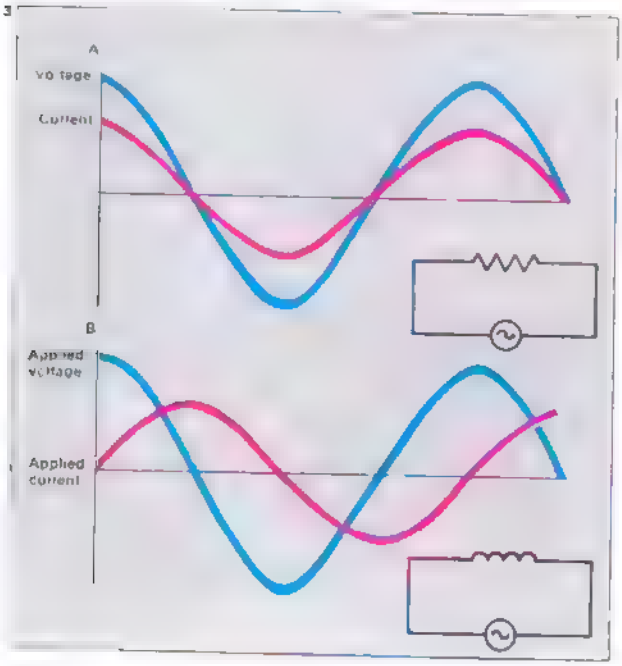
CONNECTIONS

See also

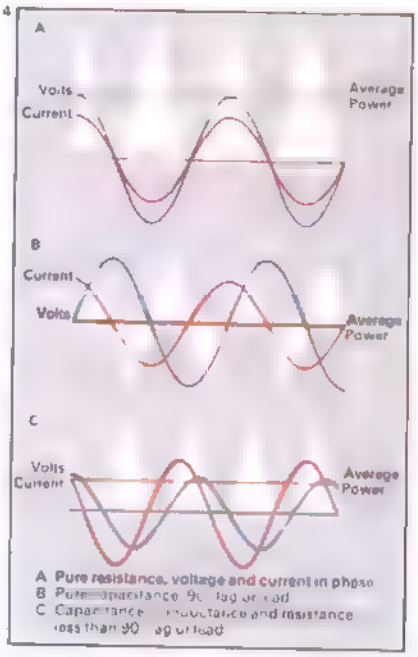
1 An inductor [A] is a circuit component consisting of a coil of wire When current flows through a coil a magnetic field is set up whose lines of magnetic flux thread through the coil Their number and distribution depend on the design of the coil As the field changes in strength with the changing current flux lines increase or decrease, cutting the windings of the coil This is the principle of a generator and an emf (electromotive force) is generated in such a way as to oppose current changes The effectiveness of an inductor can be changed by screwing a threaded iron core into or out of the coil [B]



2 A capacitor or condenser [A] is a component used primarily in alternating current circuits (particularly in electronic applications) The two parallel plates alternately store electrons as the current varies. The capacitor holds the electrons in balance and releases them at the same rate as the supply current, but out of phase with it Shown here are (top) a variable condenser consisting of parallel metal plates separated by air gaps, commonly used as the tuning control in a radio set and an electrolytic condenser (bottom) consisting of a roll of aluminium foil. In [B] a large industrial condenser is shown



3 In an AC circuit that has only resistance in it [A], the voltage applied and the current flow are exactly in phase, that is, their maxima, minima and zero points always occur at the same instant, this is true no matter how quickly the voltage fluctuates. With only inductance in the circuit, the voltage and current are out of phase. In an inductive circuit [B], current is said to lag the voltage, or the voltage lead the current In a capacitive circuit, the reverse applies – the current leads the voltage. The inductor and capacitor (unlike the resistor) store energy and release it out of phase with the input – something like flywheels on steam engines



4 Power in an AC circuit is the instantaneous product of voltage and current averaged over a fixed period In a resistive circuit (one that has resistance only) the voltage and current are in phase [A] and the power dissipated is given by the formula VI (voltage × current) In a purely capacitive circuit (one that has capacitance only) the current and voltage are 90° out of phase and the circuit returns as much power as it absorbs [B] This also applies to a purely inductive circuit (one that has inductance only) There is, however, always some resistance present, the phase angle Φ is no longer 90°, and some power is absorbed according to VIcosΦ [C]

vice versa, because each needs the current at different times in the cycle It is a repeated borrowing and lending action

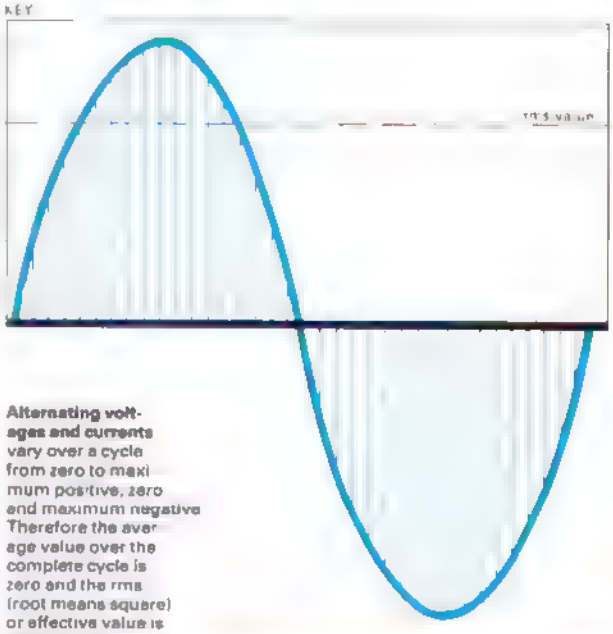
In a normal circuit some resistance R is always present. The phase angle Φ between the voltage and current depends on the inductance, capacitance, resistance and frequency Resonance can still occur when the phase lags and leads cancel out

The actual value of the current taken by a capacitor or inductor depends on three factors: the voltage, the frequency and the value (size) of the capacitor or inductor The higher the capacitance the higher the current, but for an inductor the current is smaller as the inductance rises. Current or voltage vectors may be added or subtracted using special mathematics to give a resultant vector The process is similar to that applied to mechanical vectors. Pulling a barge with ropes at an angle from each side of a canal, for instance gives a resultant forward motion

Root mean square values

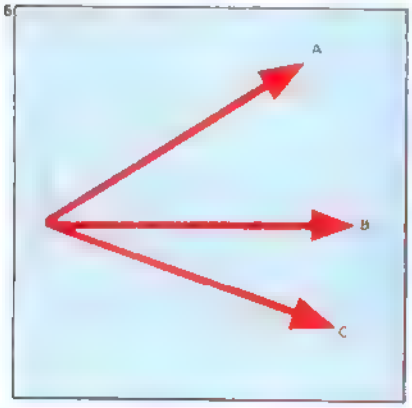
In an AC circuit the magnitude of the alternating voltage or current is defined as the rms (root mean square, or "effective") value It is used because the average value is zero, since during any short interval of time the number of half cycles in one direction equals those in the other. The rms value can be derived using simple mathematics and is that quoted in all descriptions of electrical equipment On an electric iron, a plate reading "230V, 2A" refers to rms values. The voltage and current in this case vary constantly in the form of a sine wave, reaching +325V and +2.828A – both 50 times every second

All electrical appliances that are essentially resistive in structure – such as incandescent lamps, heaters and irons – operate perfectly well in either alternating or direct current circuits (provided they are of the same voltage). But equipment that depends on inductive or capacitive properties, such as some motors, transformers and fluorescent lamps, can operate only with alternating current. Alternating current is preferred for the domestic electricity supply because it can be transmitted efficiently and easily from the power station to the domestic consumer and, is safer when it is switched on and off

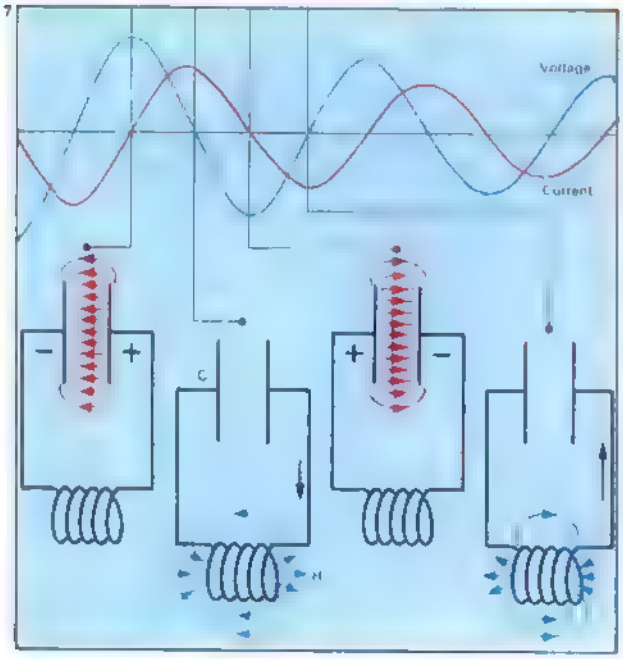


Alternating voltages and currents vary over a cycle from zero to maximum positive, zero and maximum negative Therefore the average value over the complete cycle is zero and the rms (root means square) or effective value is used as a measure.

5 A vector diagram shows the relationship between the three branches of current in a capacitance, resistance and inductance Three currents and their phase relationships with the applied voltage are depicted, by convention, by lines whose lengths represent the various values of current (they are vectors) A is said to lead, B to be in phase and C to lag behind the voltage



6 Electric clocks connected to the AC network depend on the system frequency for keeping accurate time. This is why the kitchen clock [A] always shows the same time as a master clock in the power station [B] A reduction in frequency makes the clock read slow and vice versa Because many essential timing devices depend on the mains frequency, it is maintained at a constant value (normally 50Hz in Europe) Any fall in frequency (when demands are heavy) can be gradually made up by adjusting the frequency The frequency rarely varies by more than 4 per cent



7 A circuit with inductance [H], resistance and capacitance [C] may contain some voltages out of phase with the common current It is possible for those of leading components to cancel out those of lagging components, when the circuit is said to be resonant Resonant circuits are used in radio sets

8 The power given by an AC generator depends on its voltage, current output and frequency The power to weight ratio becomes more favourable as the frequency increases, which is why aircraft generators operate at 400Hz Here an auxiliary generator is being used to power electrical circuits in an aircraft while its engines are idle

Semiconductors

Metals such as copper and aluminium are good conductors of electricity. Glass, rubber and most plastics are non-conductors, or insulators. But there are some materials, such as germanium and silicon, which are neither good nor bad conductors, and they are called semiconductors. They are used for making transistors and other solid-state devices

Current carriers

The atoms in a semiconductor easily lose one of their electrons, allowing another from a nearby atom to replace it. Although this electron exchange process goes on, the overall charge of the material is nil; in other words, it is electrically neutral. But by adding some different atoms in the form of a slight impurity, for example with one more electron per atom than the atoms of the material itself, an entirely new material is created. Just one of these new atoms (such as phosphorus, arsenic or antimony) for every one hundred thousand million germanium or silicon atoms can make a semiconductor called an *n*-type material, in which a few extra electrons are available for carrying current.

The opposite situation can be produced, making a material deficient in electrons, by adding atoms with one fewer electron per atom than those of the original material. In this case, aluminium, gallium or indium in the same remarkably small proportion is added to the germanium or silicon to produce *p*-type material [Key]. In both *n*-type and *p*-type materials, the electrons involved in the creation of the particular type of semiconductor are known as the valence electrons – the ones in the outer shell of the atom

In *n*-type materials, the surplus electrons provide the means for current flow, whereas in *p*-type surplus "holes" are created for the electrons to settle into. And as a hole exerts forces of attraction on the surrounding electrons, it can be thought of as if it were a positively charged particle. The most numerous – electrons or holes – are called majority carriers, or current carriers, in contrast to the minority carriers which are the few residual electrons or holes [1]

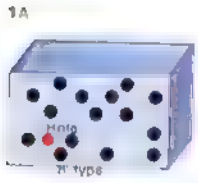
The simplest form of semiconductor device, a *p-n* junction diode [3], is made by joining pieces of opposite types of semiconductor material together, attaching a wire to each and enclosing the combination in a metal or plastic shield, with the leads protruding. By connecting a battery, so that the positive terminal is connected to the *n*-type material, a very small current flows consisting of minority carriers only. But if the battery is reversed [4], there is a large current flow, because it consists of majority carriers

Forming a transistor

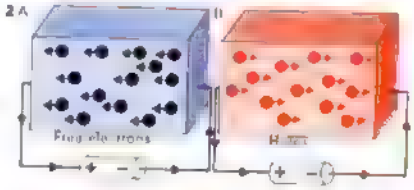
When a layer of one type of semiconductor material is sandwiched between two layers of the opposite type, a conventional two-junction, three-layer semiconductor device is formed, known as the junction transistor [5]. This arrangement can be used to form either a *p-n-p* or an *n-p-n* device. Apart from voltage polarities, both can be connected in a circuit to provide current amplifying devices. The voltages in each instance are low – for example, a voltage of 6 volts DC between the collector and base of an *n-p-n* transistor

When the base/emitter voltage is increased from, say, 600 to 620 millivolts, the collector current might increase from 0.995

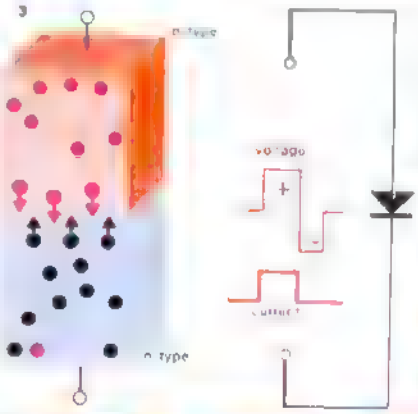
882 SEMICONDUCTORS
CONNECTIONS
See also



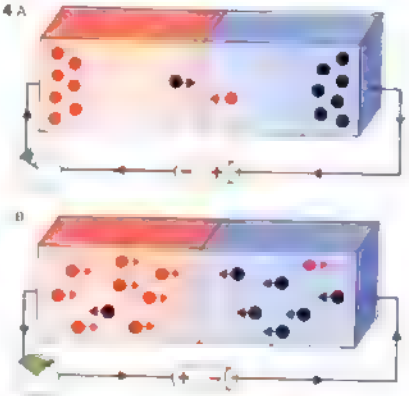
1 Electrons and holes deliberately introduced into the intrinsic (pure basic) material are seen here to be much more numerous than the carriers (holes [A] or electrons [B]) made by thermal activity



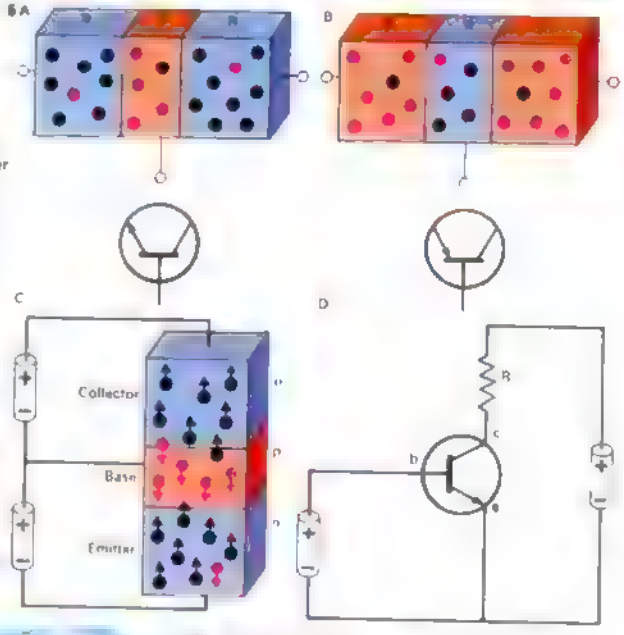
2 In an *n*-type material [A], electrons are attracted by the oppositely charged (positive) terminal of the battery. But for the fact that electrons are also flowing from the negative terminal into the other end of the material, it would be left with a net positive charge. With *p*-type material [B] a similar action ensues, but initial attraction is between positive holes and the negative terminal. The invention of the transistor in 1948 revolutionized the world of electronics. Before this no one had ever considered how vitally important semiconductors would become to the electronics industry.



3 Joining together two different semiconductor materials (*n*-type and *p*-type) causes carriers to start drifting across the junction area. As soon as a few holes and electrons have crossed the junction, they make a thin section of each material oppositely charged from the rest. A barrier which is free from carriers is therefore produced. The combination acts as a semiconductor diode

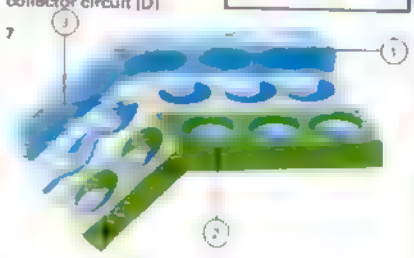


4 Wiring the terminals of a battery across a piece of joined *n*-type and *p*-type semiconductor material, in which a barrier free from carriers has been produced, simply widens the barrier if the positive pole of the battery is connected to the *n*-type [A]. Little current flows. But if the battery is reversed, forward bias is created together with the breakdown of the barrier and a large current flows [B]



5 A transistor is a sandwich of *n* and *p* type material, like two diodes back to back, [A] or [B]. In the *n-p-n* device [A, C] electrons readily flow from the *n*-type emitter to the *p*-type base under the influence of a forward bias voltage. The base is doped to give relatively few holes, so the base current is small. Most electrons are attracted across the thin base, by the reverse bias on the second diode, into the second *n*-type area called the collector. The large flow of electrons from emitter to collector is proportional to the small flow of holes from base to emitter. This makes a current amplifier. A voltage amplifier can be made by putting a resistance R in the collector circuit [D]

6 Most alloy junction semiconductor devices were of the *p-n-p* type with germanium as the *n*-type base material in the form of a wafer. A pellet of "impurity" material, such as indium, was placed on each side and heated. The indium melted, and after recrystallization, *p*-type areas formed. In recent devices, and in integrated circuits, gaseous boron or phosphorous is diffused into wafers of hot silicon to make layers of *p*- or *n*-type material



7 Some electrons in a metal are free to flow, provided that losses are made up by a reservoir, such as a battery [1]. An insulator is like an array of ponds – the potential walls around each atom (from the positive nuclei) are too high for electrons to escape – unless the system is "tilted" by a very strong electric field [2]. In semiconductors the potential walls are not so far above the electron surface. The tilt from a small electric field permits electrons to jump from one atom to the next [3]

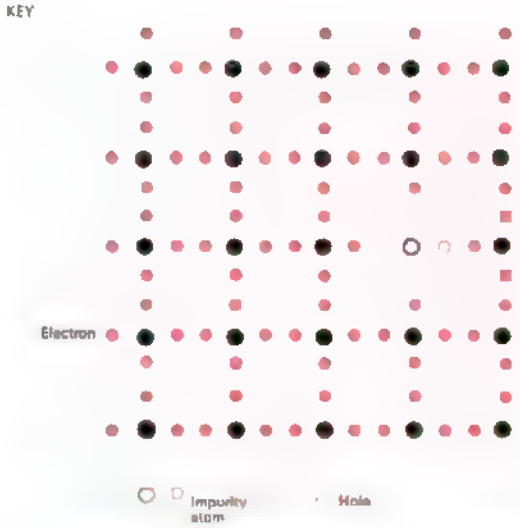
to 1 990 milliamps (mA), whereas the corresponding base current would probably increase only from 0.005 to 0.010mA. Therefore the gain of 0.005mA in the base current has caused a gain of 0 995mA in the collector current, a current amplification of 200 times.

The earliest transistors were of the point contact type, but this method of manufacture was quickly replaced by the alloy junction [6]. It involves the use of heat to form two regions of *p*-type material in an *n*-type germanium wafer. The resulting device is a *p-n-p* semiconductor, although it can handle only low currents.

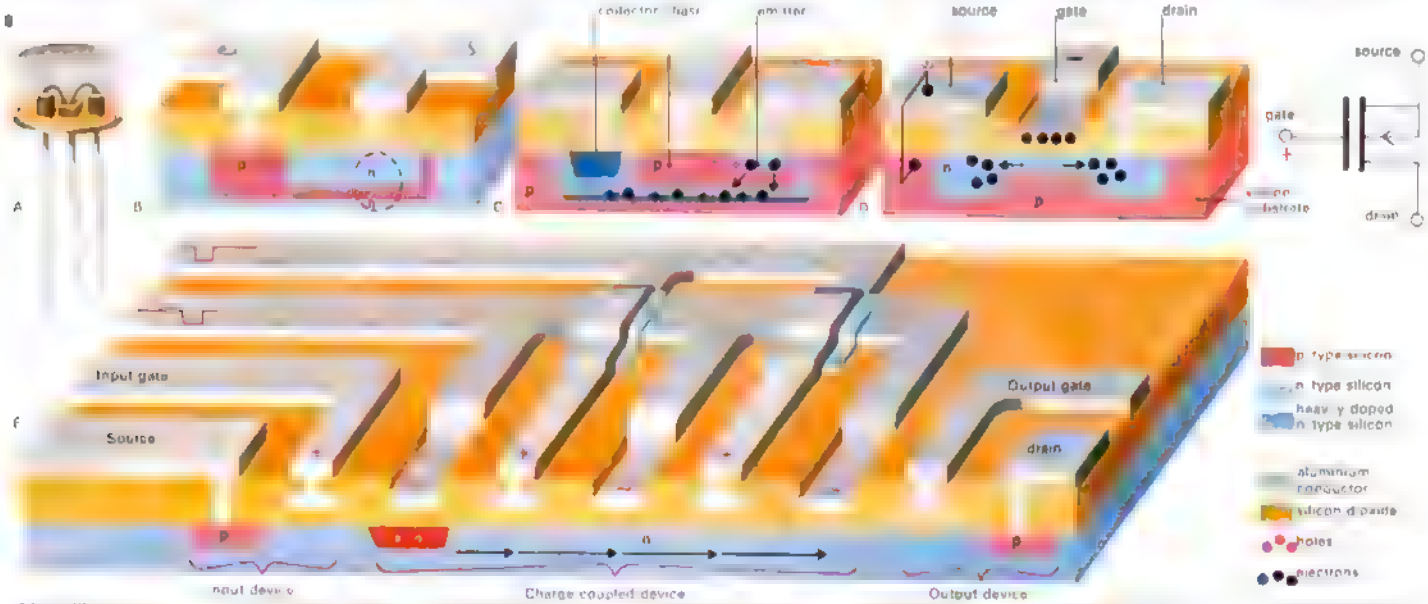
Integrated circuits

A more recent method of manufacturing semi-conductors involves the forming of p- and n-type silicon layers on the surface of a silicon wafer, sliced from a single large crystal. Many layers can be added, the atoms of each fresh layer matching up with those on the surface of the original crystal so that they behave, electrically, as if they had always been part of the crystal. The n- or p-type layers may be coated with silicon oxide for protection and insulation, or covered with metal to make contact between devices on the same chip. The shape and size of the surface deposited is controlled by masks, produced photographically or by electron beams, and circuit elements can be made as small as it is possible to make masks – typically 10^{-4}cm for a mask made photographically.

These advances in technology have generated a host of new microelectronic devices. For example, the MOSFET – Metal Oxide-Semiconductor Field Effect Transistor [8] – which can be packed more densely on a silicon chip than the conventional bipolar transistor. This allows the fabrication of quite powerful computers on a single silicon chip [10B]. Some devices have unique properties that cannot be duplicated by discrete components. In the charge couple device [8D], a row of gates act like a "bucket brigade", storing packets of charge under alternate gates to be handed on when required. Charge produced by the absorption of light can be used to drive a solid-state TV camera.



If a "foreign" or impurity atom with three electrons in its outer shell enters an array of atoms that have four electrons in their outer shells a "hole" is created in the lattice. This hole acts as an attraction to the surrounding electrons, it can be considered as a virtual positive charge.



8 In a silicon transistor [A] shown in cross-section [B], the important part, the *n-p-n* junction, is below the surface, protected by silicon oxide [1]. Both *n*-type emitter and *p*-type base were formed by diffusion through holes etched into the oxide. These were then filled in with more oxide. The chip itself is used as a collector. The base is thin so that current does not take long to diffuse across it, enabling the transistor to turn on in only 10 $^{-8}$ seconds. An integrated circuit [C] may contain many of these bipolar transistors on the surface of a single silicon chip. It is necessary to isolate them electrically or they will interfere with each other. By fabricating the *n-p-n* device on a *p*-type substrate, the *n*-type collector and *p*-type substrate act as a reverse-biased diode which, with very high resistance, achieves isolation. Field Effect Transistors (FET's), made with Metal, Oxide and Semiconductors (hence MOS device) use only one type of charge carrier, electrons or holes. In this n-MOS transistor [D] two islands of *n*-type silicon, called the source and the drain, are separated by a narrow channel which normally conducts. A negative potential applied to the gate narrows the channel and reduces the current. A charge-coupled device [E] is like a MOS device with a long string of gates between the source and the drain. A charge packet of holes can be retained by applying a steady negative voltage to a gate. If that negative voltage is reduced and, simultaneously, the next gate in the line is made more negative, the charge moves under the second gate. "Bits" of information, stored as packets of charge, can be passed along the device by applying pulses in sequence to alternate gates.

9 Photographic methods are used in combination with diffusion and etching techniques to produce large numbers of identical circuits on a wafer cut from a single crystal of silicon [A]. An array of fine probes, lined up with the aid of a microscope, is used to test a circuit [B]. The wafer is then sliced up and each chip is individually mounted, connected to terminal pins and protectively sealed against dust and damp.

10 Integrated circuits make feasible the miniaturization of all kinds of electronic devices. Such circuits are used in this electronic watch [A] and in this microcomputer [B] which contains a central processor, for performing logical and arithmetical operations, and a memory. Instructions and data are fed in through an input device such as a keyboard, output can be displayed on a TV screen or may be used to control a machine or robot.

Basic electronic principles

Everything in electronics begins with the electrons that are part of every atom. Scientists have painstakingly built up the modern picture of the atom, but no one has ever seen one because they are so small that even the most powerful electron microscopes have difficulty revealing them. Even smaller are the minute, negatively charged electrons, which can be thought of as orbiting at a distance round the atom's central nucleus where most of its mass is concentrated

The movement of electrons

Although atoms are normally neutral they can acquire an extra electron and so become negatively charged, or lose an electron and become positively charged. It is this ability of certain atoms easily to "lose" electrons that enables a stream of free electrons – an electric current – to flow in a conductor [1]. By using a battery, or a generator, a surplus of electrons can be provided at one terminal and a deficit at the other, to produce an electromotive force (emf). If a conductor is connected between these terminals the emf causes electrons to flow (or rather "drift" – the rate is seldom more than 2cm [0.75in] a minute) from the "surplus" terminal (negative) to the "deficit" terminal (positive) This is opposite to the adopted convention that assumes that electric current flows from positive to negative. Unfortunately this convention was firmly established before anyone knew anything about electrons and has been allowed to remain ever since

In electronic circuits, conductors (in the form of wires or thin copper strips on an insulating material such as paxolin) act as paths for the free flow of electrons from one part of a circuit to another. But elements are needed to control the flow, to allow precise currents of electrons to pass through various circuit components such as valves and transistors. These elements are known as resistors [2], and are available in a wide range of values from a fraction of an ohm (the unit of resistance) up to tens of millions of ohms

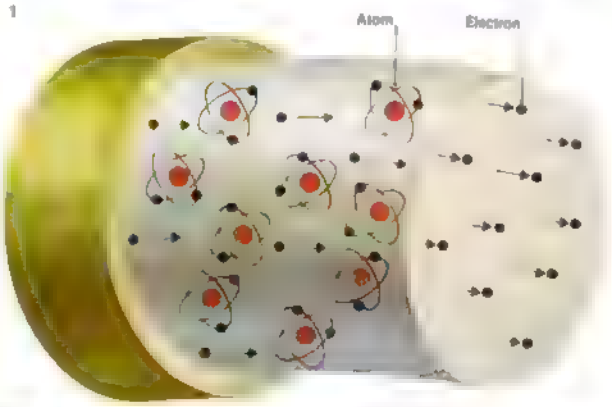
Valves and their components

The diode valve is the simplest form of vacuum tube [3A] and can change an alternating current (such as that at the mains) into a series of pulses – direct current – by a process known as rectification A diode with a single anode produces half-wave rectification [3B]. The efficiency of the process is improved by full-wave rectification [3F]. To obtain a direct current that has virtually no pulses in it at all the pulsating current can be fed into additional circuit elements, such as capacitors and chokes, which "smooth" it

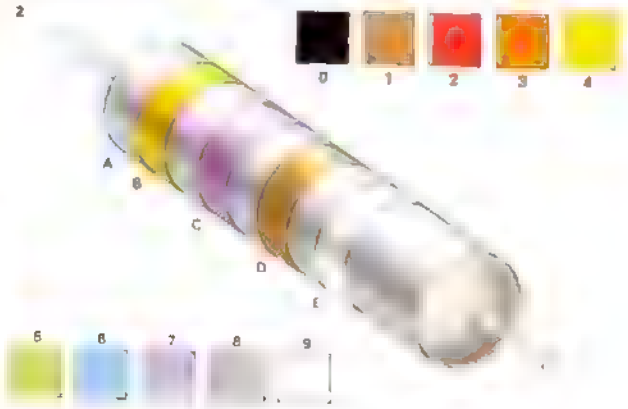
Nowadays valve diodes are usually rejected in favour of their solid-state semiconductor equivalents. Virtually all modern domestic circuitry is made up of solid-state components, apart from a few television sets that still employ one or two valves as well as semiconductors

Although the solid-state diode [3C] is much smaller than the vacuum tube equivalent it performs exactly the same rectifying function when used in a similar type of circuit [3D] But it does not have a filament (heater) and so a semiconductor diode does not consume a large amount of power. It thus makes cooling the equipment unneccessary as well as enabling the size and cost of any associated transformer to be reduced

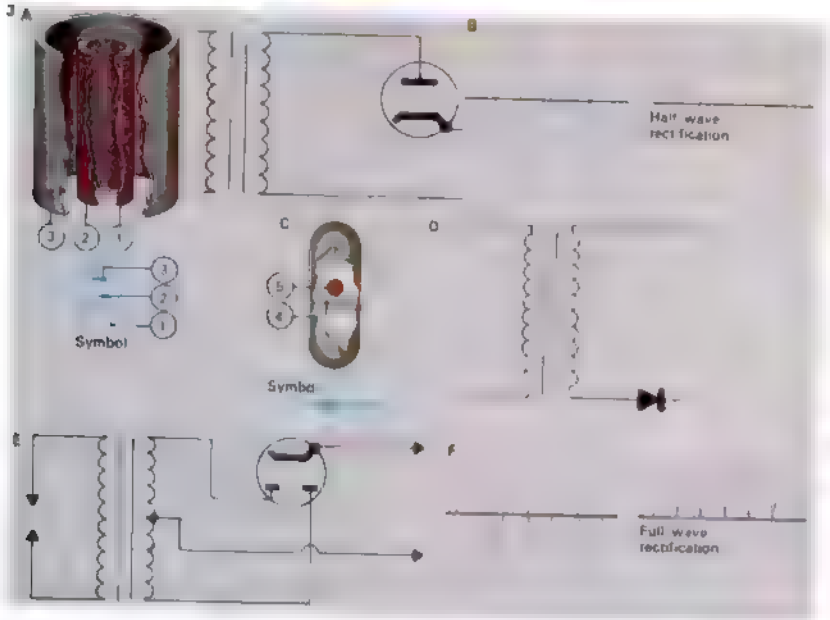
CONNECTIONS
See also



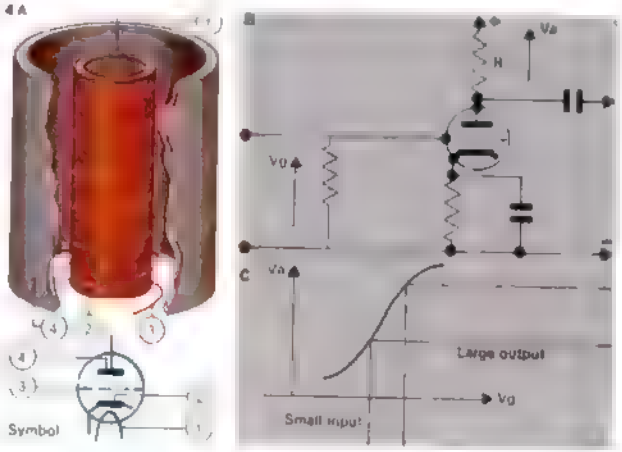
1 By applying an electromotive force (emf) to the ends of a wire made of a metal such as copper a flow of electrons can be maintained These electrons, not associated with particular atoms, are free to drift independently of each other through out the metal Under the influence of an emf, they flow, on average, in one direction and constitute an electric current



2 Carbon resistors usually have their resistance value in ohms (ranging up to millions of ohms) marked on them in an internationally recognized colour code. This resistor has a value of 470 ohms with a tolerance (accuracy) of 10% The key to this simple code is as follows: end ring [A] has no significance, ring B gives the first figure; ring C gives the second figure; ring D gives the numbers of noughts to be added to the digits in B and C. If there is a fourth ring E it indicates tolerance silver 10%, gold 5%. An absence of a fourth ring would indicate a tolerance of 20%



3 When a voltage is applied [A] to the filament (heater) [1] of a diode valve it causes the cathode [2] to emit electrons that are instantly attracted to the anode [3] As the anode cannot emit electrons, current flow is possible only in one direction. Therefore the diode valve is ideal as a rectifier of alternating current, either as a half wave rectifier [B] or a full-wave rectifier [F] using a different circuit with a pair of anodes to produce a "double diode" valve [E] The solid state component known as a semi conductor diode, is more efficient for it does not need a heater This device [C] has a layer of *p*-type semiconductor material [4] and a bead of *n*-type [5]. This combination has a low resistance in one direction and a high resistance in the other, allowing current to flow in only one direction and therefore producing rectification A half wave rectification circuit using such a diode is shown in D



4 By adding an electrode called a control grid [3] between the cathode [2] and the anode [4] the current in the valve can be controlled by the voltage on the grid As with the diode valve, a heater [1] is essential to start current flow in the resulting triode valve [A] In practice, it is found that a small change in grid voltage results in a large change in anode current. By using this effect in the circuit [B] a varying anode current can be converted to a voltage in the resistor R – the result is a signal amplified relative to the input at the control grid [C]

By adding an extra electrode to the diode valve in 1906 the American inventor Lee De Forest (1873–1961) controlled the flow of electrons between the cathode and the anode [4A]. And by adding other basic circuit elements, in this case resistors and capacitors, the triode valve could be used as a voltage amplifier [4B]. Later, other grids were added to the triode to improve performance, especially in early radio receivers.

The transistor

Following a successful research programme directed by William Bradford Shockley (1910–) at the Bell Telephone Laboratories, Murray Hills, New Jersey, USA, the world of electronics was suddenly presented with the first solid-state three-electrode device, which was destined to end the supremacy of the electronic valve. Shockley, John Bardeen and Walter Houser Brattain were awarded the Nobel prize in physics in 1956 for work on the development of the transistor in 1948.

As a result of this work the world now enjoys the benefits of small, inexpensive portable radios that run on torch batteries [5]. Inspecting a circuit shows that transistors are even smaller than many of the other conventional components, despite the fact that these are now miniaturized. A transistor consists of a layer of one type of semiconductor material between two layers of a different type. These materials are called *p*-type and *n*-type and either of them can form the inner layer. Taking up even less room, an integrated circuit has one or more transistors and other circuit components formed within a single "chip" of semiconductor material.

An interesting feature of practical valve and transistor circuits is the use of "feedback" to improve the quality of the sound produced by an amplifier. This is a refinement to overcome the distortion of the signal that can take place between the input and the output circuits of a valve or transistor or of several such devices. Feedback may be voltage feedback or current feedback.

In small portable transistor radios, there is seldom any output transformer at all because a "single-ended" push-pull output stage is used, reducing cost and weight.

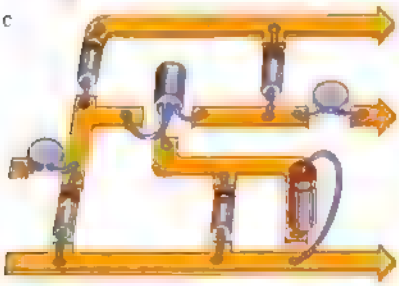
KEY

The mysteries of the atom have fascinated many artists. "Theme on Electronics" by the English sculptor Barbara Hepworth (1903–75) was created in 1957 for the Mullard Electronics Centre in London. It symbolizes the world of the electron. Every atom can be pictured with one or more electrons orbiting a central nucleus (with one or more protons and sometimes one or more neutrons as well). The electron always carries a negative charge and the central nucleus a positive one. Normally these charges balance each other out exactly so that neither the negative nor the positive charge predominates – in other words an atom is normally electrically neutral.

5A

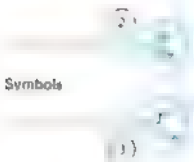
B

C

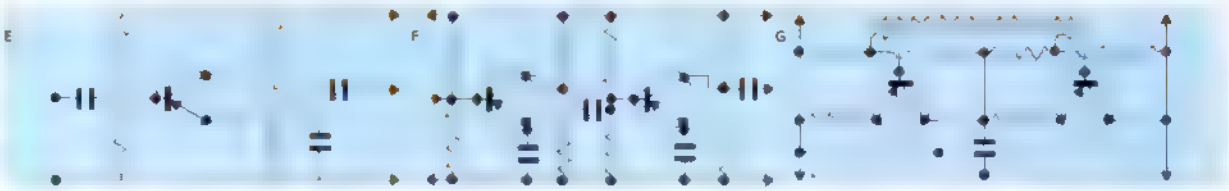


D

Symbols



E F G



5 A modern portable radio (A) owes its existence to the invention of the transistor in 1948. Without transistors such devices as this one were almost out of the question as far as mass production was concerned because valves, and most other parts were too large. Looking inside the set (B) it is possible to see how many transistors and other components can be packed into such a small space and yet leave room for the batteries and loudspeaker. Within a portable transistor radio there are several types of amplifying circuits. In a typical one-stage transistor amplifier (C), several components are grouped around the transistor, which emphasizes its small size even more. The upper wires of the various capacitors and resistors are cleverly folded back and joined onto the printed circuit board. Basically, the transistor itself (D) comprises a centre layer called the base (1) of one type of semiconductor material sandwiched between two other layers of the opposite type of semiconductor material (the collector and the emitter). The two types are *n*-type and *p*-type and the composition of the sandwich is denoted by the names of the transistor types *pnp* (2) and *npn* (3). The essential difference is in the polarities of the three terminals, and the two types of transistors are distinguished by the directions of the emitter arrows in the symbols. The circuit (E) represents the amplifier circuit (C). Increased amplification of small signals can be obtained by using several amplification stages (F). Where high output is needed a "single-ended", push-pull circuit (G) is used, each amplifying half of the signal.

What is chemistry?

Dyestuffs, drugs, synthetic fibres, photographic products, detergents, fertilizers – these are just a few examples of the very many products that have been made through chemistry. But what is chemistry? Every science looks at the world in its own special way; the basic building block with which chemistry is concerned is the atom. Chemistry deals with the properties of different atoms, the ways in which they join together to form molecules, and the interactions of various kinds of molecules with one another

The stuff of atoms

To the chemist, an atom is made up of three kinds of sub-atomic particles – the proton, neutron and electron. The only difference between a neutron and a proton is that a neutron has no electric charge, whereas a proton has a unit of positive charge. An equal unit of negative charge is carried by the electron, which is much lighter

In any atom, the protons and neutrons are packed closely together in the central nucleus. Surrounding this, but much less closely packed, are the atom's electrons. The radius of a neutral atom – that is, one in which there are as many electrons as protons – is about 10,000 times larger than the radius of its nucleus. An atom is composed largely of empty space. Because of this, it is much more likely, when two atoms collide, that their electrons will interact with one another than that the two nuclei will ever come into contact Consequently, chemists are concerned primarily with the electrons in atoms

Different kinds of atoms result from the combination of different numbers of protons, neutrons and electrons. The number of protons in an atom is its atomic number, and the total mass of all the sub-atomic particles (protons, neutrons and electrons) is its atomic weight. The simplest atom, that of hydrogen, consists of a single proton and a single electron. If a neutron is added to the nucleus of hydrogen, a different kind of atom, called deuterium, is formed. In many ways, the behaviour of hydrogen and deuterium is the same – as one might expect of two atoms each with only one electron. As a result, the hydrogen atom with no neutron (sometimes called protium) and the deuterium atom are classed by the chemist as being different isotopes of the same element There is also a third isotope of the element hydrogen. This is tritium, with two neutrons and one proton in its nucleus. But if a second proton is added to a tritium nucleus and, to balance the electric charge, another electron is placed round this nucleus, the situation is quite different. The atom shows no resemblance in its chemical behaviour to any of the hydrogen isotopes. It is an atom of an entirely different element – helium

How elements are built up

The element to which any atom belongs is defined by the number of protons in its nucleus. The number of neutrons can usually vary slightly, to give a range of isotopes. The isotopes of an element have different masses but identical chemical properties. There are many stable isotopes in nature, but among the elements with about 90 or more protons, such as uranium, the isotopes tend to be unstable and the nuclei break down to form atoms of other elements [4]. Nuclear reactors and atomic bombs depend on this instability

CONNECTIONS

See also

1

1 This iron rosette, dating from the 7th century BC, once decorated an Etruscan chariot. Techniques of working iron and, earlier, copper and bronze go back thousands of years. They allowed men to make tools and weapons, implements that advanced both agriculture and warfare. Extraction and purification of metals probably developed accidentally, but led to the founding of present day chemistry

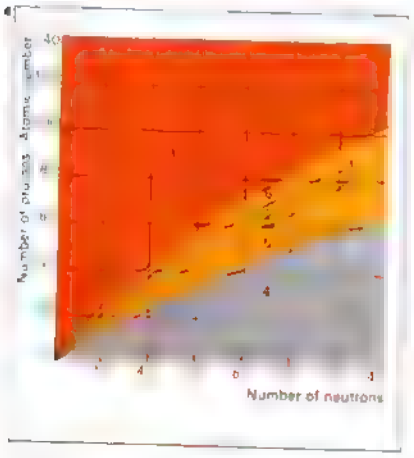
2

2 Ancient Greeks, such as Democritus, believed that matter is made from tiny particles, which they called atoms But "atomism" did not become a useful part of chemistry until John Dalton (1766–1844) proposed that atoms have different weights, and that the weight of any particular atom is constant. By 1808 Dalton had drawn up this list of symbols for the different types of atoms (elements)

3

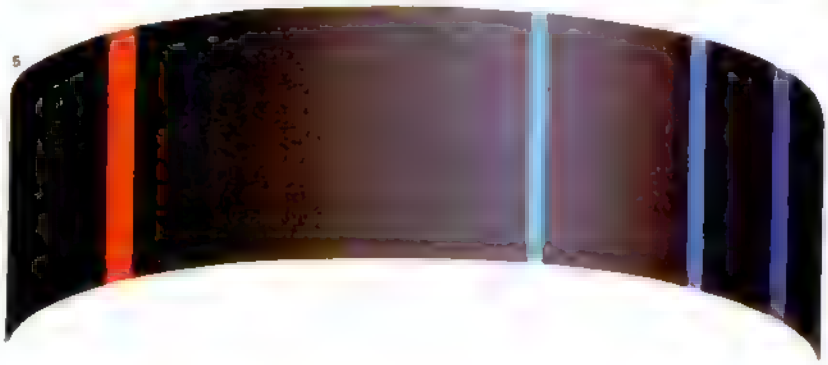
3 John Dalton, the son of a weaver, is one of chemistry's most distinguished figures as a result of his work on the atomic theory. He was born in Cumberland, England, in 1766, into a Quaker family At the age of 12 he became a school teacher He moved to Manchester in 1793 and stayed there until his death in 1844, earning his living by teaching mathematics and natural philosophy (physics) Dalton was colour-blind and in addition to the atomic theory which he developed during the first decade of the 1800s, he also investigated colour-blindness (sometimes called Daltonism) and meteorology

4



4 The nucleus of an atom is made up of protons and neutrons. Only certain combinations of these sub-atomic particles are stable [2] Atomic nuclei containing a large excess of protons [1] or of neutrons [4] are both highly unstable Beyond the stable region is a group of atoms [3] where nuclei break apart to form smaller, more stable nuclei Below [5], and above [6], the stable regions are unstable nuclei, which decay into stable elements

5



5 Light of specific wavelengths is emitted by an element such as hydrogen when it is "excited" by passing an arc discharge through it

If light of these wavelengths is passed through the vapour of the same element, it is absorbed to form an absorption spectrum.

In the 1800s men wondered why elements should have such discrete spectra, and why the spectral lines could be divided into four classes – called sharp, principal, diffuse and fundamental. Only when modern atomic theory was developed was an explanation found

As the elements build up, their nuclei are surrounded by more and more electrons which are arranged according to definite rules. The positive charge of the protons in the nucleus attracts the negatively charged electrons, whose momentum prevents them from "falling" into the nucleus.

Electron orbitals

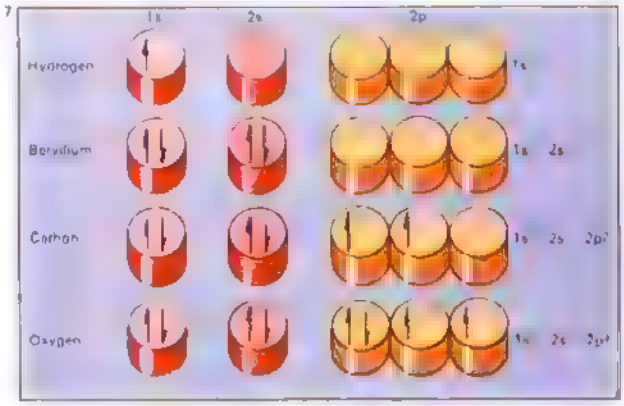
The electron in a hydrogen atom spreads out round the nucleus in a spherical shell. It is possible to assign to any point only a probability that the electron is there at an instant in time. The region round a hydrogen nucleus where there is the highest probability of finding the electron is the electron's orbital.

In 1925, the Austrian physicist Wolfgang Pauli formulated rules for electron orbitals. His major rule was that no two electrons in the same atom can be in exactly the same quantum state. The quantum state of an atom is defined by four different numbers. The first of these, known as the principal quantum number, describes the average distance between the electron and the nucleus. The second quantum number is related to the shape of the orbital, which is not always spherical, and the third determines its orientation. The final quantum number is the spin of the electron, which can only be $+\frac{1}{2}$ or $-\frac{1}{2}$. The possible values for the second and third quantum numbers depend on the value of the principal quantum number in such a way that, when it is 1, there are only two possible quantum states (and orbitals); when it is 2, there are 8, when it is 3, there are 18, and so on, according to the formula $2n^2$, where n is the principal quantum number.

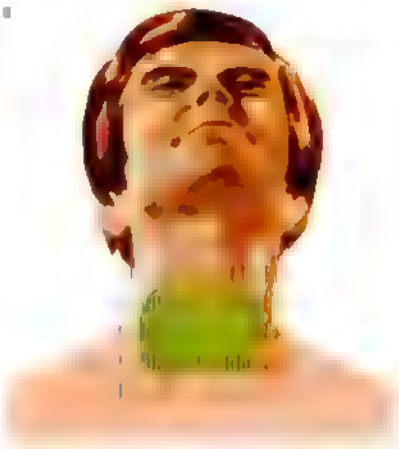
The energy of any electron depends upon the first two quantum numbers and, because its behaviour depends largely on its energy, chemists have developed a kind of shorthand for describing an electron's energy level. Each electron in an atom can be described by a number followed by a letter. The number is the same as the principal quantum number, but the values 0, 1, 2 and 3 for the second quantum number are represented (for historical reasons) by the letters s, p, d and f [7]. By knowing the numbers of electrons in different orbitals, a chemist can predict the behaviour of any particular atom.

Flasks, beakers and the simple test-tube have long symbolized the chemist's search for new elements and compounds.

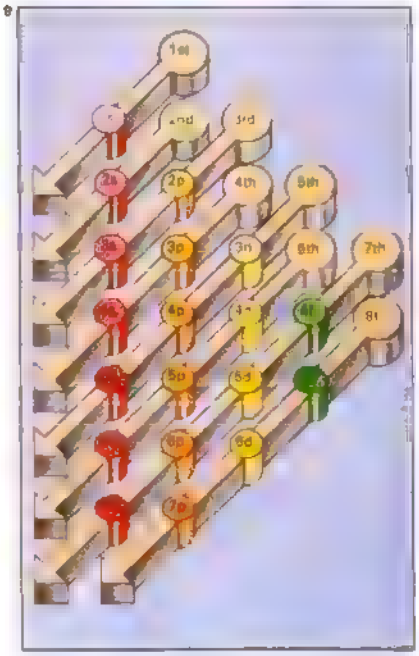
6 The isotopes of an element all contain the same number of protons but differ in neutron number. The hydrogen atom has one proton. The common form has no neutrons, the rarer form, deuterium, has one; tritium, with two neutrons, is radioactive. The stable isotope of carbon, carbon-12, has six protons and six neutrons; carbon-14, with eight neutrons, is radioactive. The ratio of these two in once-living material helps determine its age.



7 The positions of electrons in a neutral atom of any element can be shown in a "pigeon-hole" notation. The orbitals are signified by a number – the principal quantum number that indicates the average distance of the electron from the nucleus – and a letter that indicates the shape of the orbital. The letters used for identification are derived from the initial letters of the four types of spectral lines, s, p, d and f and then proceed alphabetically. Because electrons – shown by half arrows – can spin in opposite directions, each pigeon-hole can hold two electrons.



8 The breakdown of unstable isotopes of different elements can be detected and sometimes utilized. Radioactive isotopes (also called radioisotopes) have widespread applications in medicine. For example, a dose of radioactive iodine is taken up preferentially by the thyroid gland in the neck. Its presence can be detected and used to map the gland, to see if it is diseased, cancerous or less (in the one pictured here) enlarged.



9 Atoms have more and more electrons as they build up to form the heavier elements and the order in which they occupy orbitals depends on the particular binding energy. In general, the closer an electron is to the nucleus, the greater is the energy. As electrons get farther away from the nucleus, the energy relationships become more complex. So that, for example, the 4s orbital is more strongly binding than the 3d orbital, and is filled with electrons before it. The chart shows the general order for the filling of orbitals as atoms get larger. However, even this is an approximation and there are a few exceptions to it.

Classification of elements

Since man first began to purify metals from rocks thousands of years ago, he has been learning how different substances behave and trying to detect a pattern in that behaviour. But the major breakthrough in discovering the pattern of chemistry did not come until just over 100 years ago when the Russian chemist Dmitry Mendeleyev [Key] proposed his periodic system of the elements

Earlier, the French chemist Antoine Lavoisier (1743–94) had revived Robert Boyle's use of the word "element" for substances that could not be broken down into anything simpler. During the next 75 years, many new elements were discovered [4] and substances previously thought to have been elements were shown to be compounds – two or more elements combined

Atomic weights

As more elements were discovered, and more of their properties catalogued, it became clear that some elements were similar to others. Sodium and potassium, for example, first isolated in the early 1800s by Humphry Davy (1778–1829), are both soft metals that react violently with water to produce alkaline solutions. It gradually became obvious that there must be a way of tabulating the elements so that those with similar properties were grouped together

One property of elements that was being catalogued at that time was atomic weight An atom is extremely small – a toy balloon might easily hold a quadrillion of them (that is 1 followed by 24 zeros). Nevertheless each atom does have a definite mass. Most of this comes from the neutrons and protons in its nucleus. For example, a deuterium atom with one proton and one neutron in its nucleus – is almost twice as heavy as a hydrogen atom, with its one proton. An oxygen atom, with eight neutrons and eight protons, is about 16 times as heavy as a hydrogen atom

Using various analytical skills, nineteenth-century chemists gradually catalogued the comparative atomic weights of the elements with increasing accuracy. The weight of a hydrogen atom was formerly taken as 1 and the weights of other atoms related to it; atomic weights are now based on a value of 12 for carbon-12 (six neutrons and six protons), making hydrogen 1.008

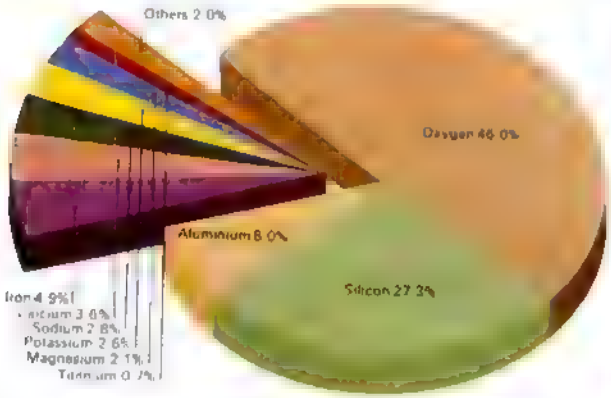
By organizing the elements in tabular form in order of increasing atomic weight, Mendeleyev produced a periodic table Unlike those proposed by other chemists, his had gaps in it. Where the difference in atomic weight between two neighbouring elements seemed exceptionally large, he assumed that there was an element of intermediate weight that had yet to be discovered Because his scheme also arranged elements in families, he was able to predict the properties of some of these undiscovered elements [6] Before the end of the nineteenth century, his assumptions were vindicated by the discovery of some of these "missing" elements and by the close agreement between Mendeleyev's predicted and their actual properties

Atomic numbers

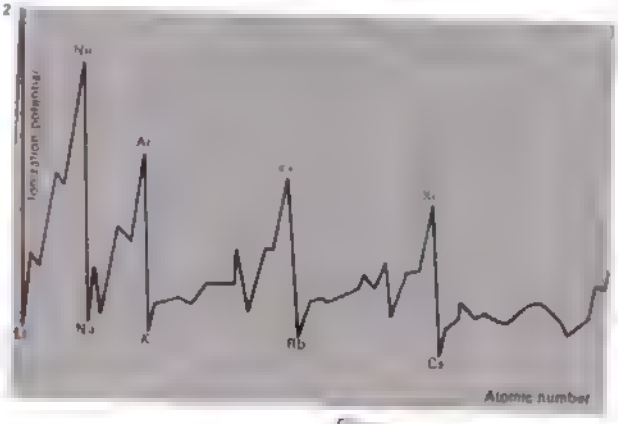
It is known that the physical basis underlying the periodic classification is not the atomic weights of the elements, but their atomic numbers – that is, the numbers of protons in their nuclei. And the regularities observed in different families of elements result from

CONNECTIONS

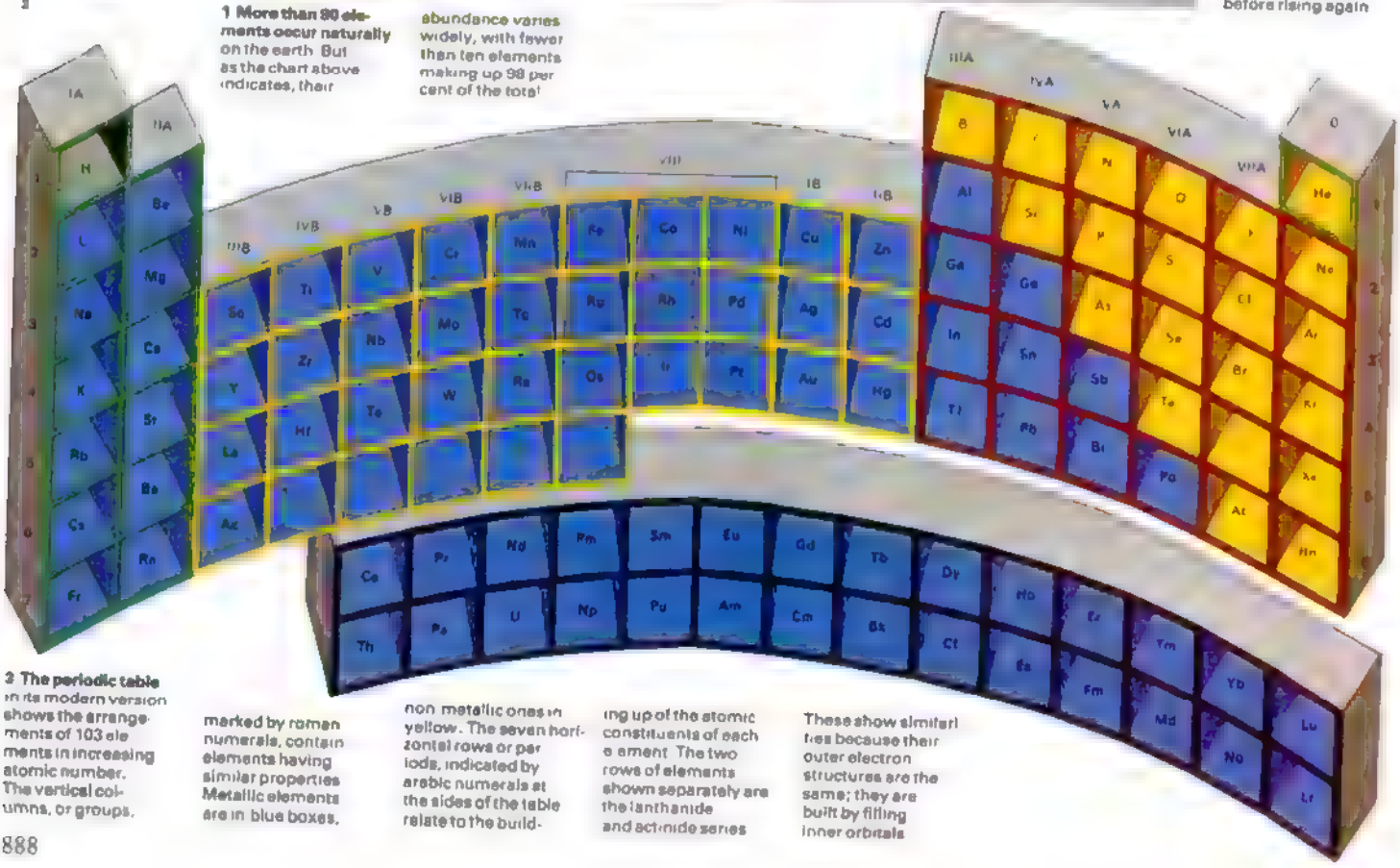
See also



1 More than 90 elements occur naturally on the earth But as the chart above indicates, their abundance varies widely, with fewer than ten elements making up 98 per cent of the total



2 Energy is needed to remove a single electron from an atom of any element The amount needed, called the first ionization potential, is different for each element When all these energies are plotted on a graph, with the elements in order of increasing atomic number along the bottom, it can be seen that they vary in a periodic manner – rising gradually and then suddenly falling, before rising again



2 The periodic table in its modern version shows the arrangements of 103 elements in increasing atomic number. The vertical columns, or groups, marked by roman numerals, contain elements having similar properties Metallic elements are in blue boxes, non metallic ones in yellow. The seven horizontal rows or periods, indicated by arabic numerals at the sides of the table relate to the building up of the atomic constituents of each element The two rows of elements shown separately are the lanthanide and actinide series These show similarities because their outer electron structures are the same; they are built by filling inner orbitals

similarities in electronic arrangement [5].

Lithium (Li) has three electrons and potassium (K) 19. According to the rules for the occupancy of orbitals (the regions in space where electrons are most likely to be found), lithium has as many electrons as are allowed (two) in the orbital with the principal quantum number 1. It also has one electron left over: this occupies the next lowest energy orbital, called 2*s*. Potassium, on the other hand, has electrons filling all possible orbitals in the first three levels, and one more. This last electron is in the 4*s* orbital. Thus, in both elements, the outermost occupied orbital has a single electron in it.

If two atoms have the same number of electrons in the outermost orbital layer, or "shell" as it is usually called, it is reasonable to expect that their chemical properties should be similar. All the elements beneath lithium in a modern periodic table [3] have only one electron in their outermost shell.

Man-made elements

For many years, it was believed that element 92 (uranium) had the heaviest atoms occurring naturally on earth. It was believed that as atomic size increases, the atoms become less stable and any atoms of heavier elements that might once have been present on earth had broken down. Since 1940, chemists in the United States and the Soviet Union have been making "transuranium" elements artificially [7]. Glenn Seaborg (1912–), who has been involved with much of this work, has used the periodic table to predict the likely properties of transuranium elements. His predictions go as far as element 168, although, for physical reasons, few of these elements would be stable for long enough to check whether or not their properties did coincide with Seaborg's predictions.

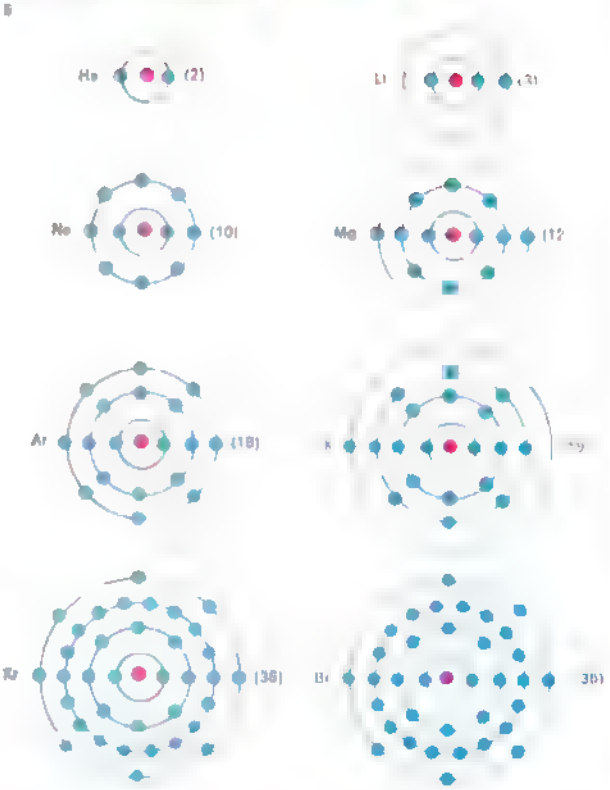
But another of his theories based on the periodic table has been proved. Because of their chemical similarities, members of the same family of elements often occur in the same minerals. Seaborg predicted that, if there were any traces of transuranic elements left on earth, they would be found in minerals rich in other elements of the same family. In 1971, he discovered naturally occurring plutonium in a sample of uranium ore.

The Russian chemist Dmitry Mendeleyev was responsible for proposing the periodic table of the elements in 1869. This major theoretical breakthrough provided the necessary classification system for making the similarities in the properties of certain elements understandable. Mendeleyev was born in Siberia in 1834 and trained as a teacher in St Petersburg. Subsequently he became professor of chemistry at St Petersburg University, a post which he held until 1890. He died in 1907.

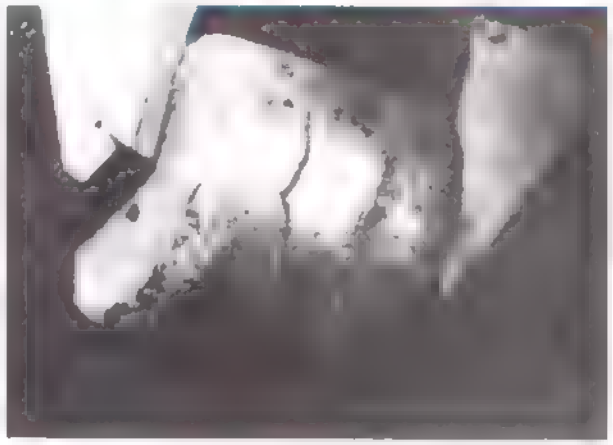
4 **The elements, their symbols and atomic numbers**

Actinium	Ac	89	Hafnium	Hf	72	Promethium	Pm	61
Aluminium	Al	13	Helium	He	2	Protactinium	Pa	91
Americium	Am	95	Holmium	Ho	67	Radium	Ra	88
Antimony	Sb	51	Hydrogen	H	1	Radon	Rn	86
Argon	Ar	18	Indium	In	49	Rhenium	Re	75
Arsenic	As	33	Iodine	I	53	Rhodium	Rh	45
Astatine	At	85	Iridium	Ir	77	Rubidium	Rb	37
Barium	Ba	56	Iron	Fe	26	Ruthenium	Ru	44
Berkelium	Bk	97	Krypton	Kr	36	Samarium	Sm	62
Beryllium	Be	4	Lanthanum	La	57	Scandium	Sc	21
Bismuth	Bi	83	Lawrencium	Lr	103	Selenium	Se	34
Boron	B	5	Lead	Pb	82	Silicon	Si	14
Bromine	Br	35	Lithium	Li	3	Silver	Ag	47
Cadmium	Cd	48	Lutetium	Lu	71	Sodium	Na	11
Caesium	Cs	55	Magnesium	Mg	12	Strontium	Sr	38
Calcium	Ca	20	Manganese	Mn	25	Sulphur	S	16
Californium	Cf	98	Mendelevium	Md	101	Tantalum	Ta	73
Carbon	C	6	Mercury	Hg	80	Technetium	Tc	43
Cerium	Ce	58	Molybdenum	Mo	42	Tellurium	Te	52
Chlorine	Cl	17	Neodymium	Nd	60	Terbium	Tb	65
Chromium	Cr	24	Neon	Ne	10	Thallium	Tl	81
Cobalt	Co	27	Neptunium	Np	93	Thorium	Th	90
Copper	Cu	29	Nickel	Ni	28	Thulium	Tm	69
Curium	Cm	96	Niobium	Nb	41	Tin	Sn	50
Dysprosium	Dy	66	Nitrogen	N	7	Titanium	Ti	22
Einsteinium	Es	99	Nobelium	No	102	Tungsten	W	74
Erbium	Er	68	Osmium	Os	76	Uranium	U	92
Europium	Eu	63	Oxygen	O	8	Vanadium	V	23
Fermium	Fm	100	Palladium	Pd	46	Xenon	Xe	54
Fluorine	F	9	Phosphorus	P	15	Ytterbium	Yb	70
Francium	Fr	87	Platinum	Pt	78	Yttrium	Y	39
Gadolinium	Gd	64	Plutonium	Pu	94	Zinc	Zn	30
Gallium	Ga	31	Polonium	Po	84	Zirconium	Zr	40
Germanium	Ge	32	Potassium	K	19			
Gold	Au	79	Praseodymium	Pr	59			

5 **The periodicity of the elements** is now known to be due to their electronic structures. The electrons in any atom can be envisaged as occupying "shells" each of which can hold only a certain number of electrons. For the first three shells, the numbers are 2, 8 and 18. If the outermost shell is either completely filled, or holds 8 electrons, the element is chemically unreactive: the gases helium (He), neon (Ne), argon (Ar) and krypton (Kr) all fulfil this condition. Lithium (Li) and potassium (K) are similar because they both have one outer electron, and both differ chemically from magnesium (Mg) and bromine (Br).



6 **Germanium**, a metal "grown" in this form for making transistors, was unknown when Mendeleyev proposed his periodic table. However, in 1871 he predicted the existence of an element with the properties of germanium and called it eka-silicium, from silicium (the old name for silicon) and the Sanskrit word *eka*, meaning "one". Germanium was discovered in 1886, 15 years after he had made his prediction.



7 **The first samples of element 94**, plutonium, were produced in 1940 in an American laboratory. One of 11 "transuranium" elements that have now definitely been synthesized, it is used as a fuel in some types of nuclear reactors.

Survey of groups of elements

Atoms of all elements consist of a central nucleus surrounded by a "cloud" containing one or more electrons. The electrons can be thought of as occupying a series of well-defined shells. The behaviour of a particular element depends largely on the number of electrons in its outermost shells. Other factors, such as the total number of electron shells, also play a part in determining behaviour but it is the dominance of the outer electron configuration that underlies the periodic law and justifies the grouping of the elements into groups or families.

The *s*-elements and their reactions

Each electron shell is made up of various volumes in space called orbitals, known as *s*-, *p*-, *d*- and *f*-orbitals, and each is at a higher energy than the one below it. Those elements in which the outermost shell can have only one or two electrons can be grouped together as the "*s*-elements" (because it is only the *s*-orbital that is occupied in the outer shell). These are (in addition to hydrogen and helium) lithium, beryllium and the elements directly below them in the periodic table [1]. All but helium readily form positive ions by the loss of their outer electrons and they are therefore mostly found as components of ionic compounds, commonly called salts. Many common substances contain these elements – for example, soda (sodium, Na), potash (potassium, K), gypsum (calcium, Ca) and carnallite (magnesium, Mg).

The *s*-elements with only one electron are more reactive than those with two. Thus if dropped on to water, sodium reacts so violently that it catches fire; if magnesium is dropped into hot water the reaction (release of hydrogen) gives off light, but is less violent.

The energy levels of electrons in the heavier elements are slightly complicated and, for this reason, many elements have a filled *s*-level but only a partly filled *d*-level below it. The part of the table ten elements wide beginning with scandium (Sc) includes the "transition" elements in which a *d*-level is successively filled. The outermost electron shell of these elements has an *s*-configuration, but it is the underlying layers of *d*-electrons that determines the element's chemical behaviour. The lanthanide and actinide series form further sub-groups: in their cases an *f*-group on a lower level begins to fill while the levels beyond contain one *d*-electron and two *s*-electrons.

The *d*-elements: "rare earths" and metals

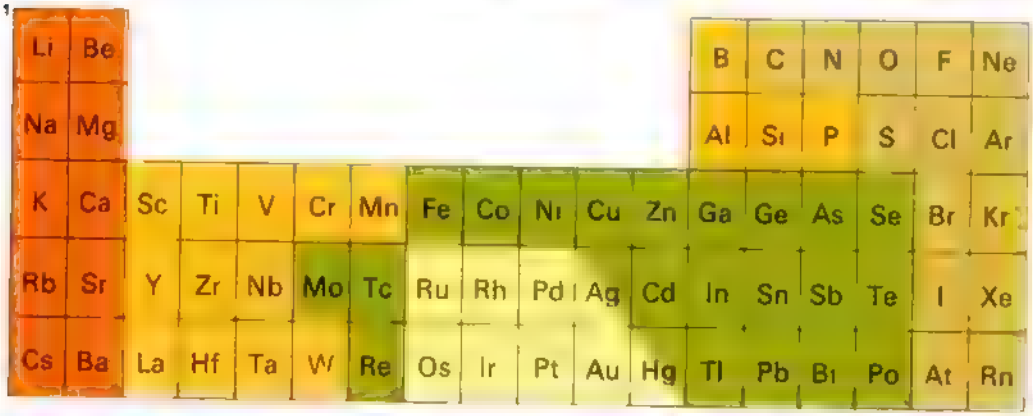
For the elements scandium, yttrium (Y) and lanthanum (La), as well as the entire 14-element lanthanide series, the chemical behaviour is very strongly influenced by the presence of the single *d*-electron. All these elements tend to form positive ions by the loss of this electron and the two *s*-electrons giving ions with a charge of +3. The elements are all fairly reactive and all are rare, but some of these "rare earths" have found commercial uses, for example in the manufacture of specially tinted spectacles.

The other sub-series, the actinides, is of greater importance as it contains the nuclear reactor fuel elements uranium and plutonium. But the importance of these elements is based on their nuclear instability rather than their chemical properties.

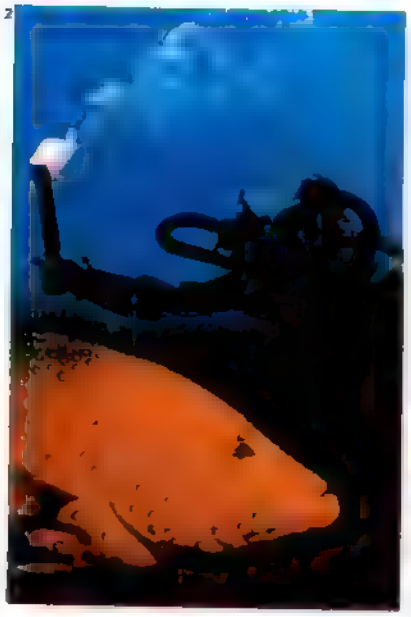
The other transition elements [4] are also important because they are all metals and

CONNECTIONS

See also



1 Elements are won from minerals in different ways according to their position on the periodic table. In the red group are reactive metals usually extracted by electrolysis. In the orange group, the elements frequently occur in ionic compounds, often combined with oxygen. These elements are often prepared by electrolysis. The third group (coloured green) is commonly found as sulphides and the elements are obtained by roasting the sulphide and reducing the resultant oxides. Group four (coloured yellow) are relatively unreactive elements, found free or as compounds that give the element when they are heated to a certain temperature. The fifth group (light brown) are non-metals that occur free or as negative ions that can be converted to elements by electrolysis.



2 Magnesium is one of the reactive *s*-elements. It burns with an intense white flame, which made it useful in early "flash" photography. It is now used in flames that burn under water.

3 Tungsten was discovered in 1783 by the brothers Fausto and Don Juan d'Elhuyar. A fine wire made from tungsten becomes white hot when carrying an electric current, as in an electric lamp.

4 Cobalt, one of the transition metals, is the basis of various blue pigments. It is also an important constituent of some biological molecules. However, a most important use in recent years has been in the medical "cobalt bomb". A radioactive isotope of cobalt (cobalt-60) which gives off high-energy gamma rays, is used to direct this cell-destroying radiation at tumour sites in humans. It is used extensively in hospitals to treat cancers and arrest their growth.

5 A compound exhibits a specific colour when it absorbs white light – a mixture of all colours – by selectively reflecting a few wavelengths. Many salts of transition metals such as iron and nickel, are coloured. The exact colour depends on which other atoms are associated with the metal. Thus chromium compounds have for centuries been used to provide painters with yellow pigments. Vincent van Gogh's "Sunflowers", shown here, is a good example of their use.

many of them have large-scale industrial uses. All of these elements – except copper silver and gold [6] – have two electrons in an outer *s*-orbital and between two and ten electrons in the underlying, but more strongly bonding *d*-orbital. Silver, copper and gold could be expected to have two *s*-electrons and nine *d*-electrons. But, because the complete filling of a *d*-shell increases stability, they have only one *s*-electron and a full complement of ten *d*-electrons.

The transition elements are characterized by the ability to form several different ions because of the complex behaviour of electrons in the *d*-orbitals. Thus iron is found in ionic compounds as both Fe^{2+} (ferrous, with two electrons lost) and Fe^{3+} (ferric, with three electrons lost from the atoms).

The *p*-elements and their grouping

To the right of the transition elements in the periodic table are the *p*-elements. In the groups headed by boron, carbon, nitrogen, oxygen, fluorine and helium, it is the three *p*-orbitals (capable of holding a maximum of six electrons) that are the most important in determining chemical behaviour. From left to right through this group the *p*-orbitals become increasingly filled, until the "completed octet" (two *s*- and six *p*-electrons) of the noble gases – the group that includes neon. These elements are unreactive, and only recently have chemists made compounds including them.

The group headed by fluorine – the halogens – all need only one electron to complete the octet and they readily do this to form ions carrying a single negative charge. The oxygen group can form double-negative charged ions, but tend more to link with other atoms through covalent (non-ionic) bonds. The tendency towards covalent bonding is even more marked in the groups headed by nitrogen and carbon. The group headed by boron, where there are two *s*-electrons and one *p*-electron, like the group headed by scandium, shows more of a tendency to form ions. Aluminium, for example, readily loses its three outer electrons to form Al^{3+}, but aluminium chloride ($AlCl_3$) is not an ionic compound. The bonds joining the chlorine and aluminium atoms are partly ionic and partly covalent.

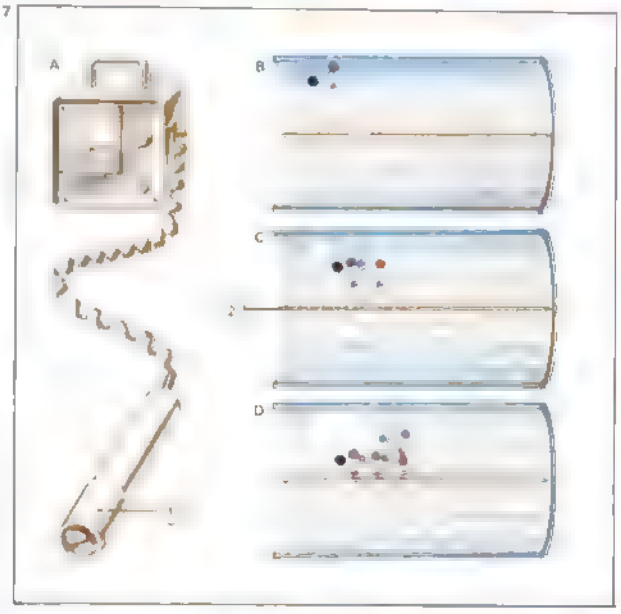
KEY

Elements such as iron and lead have been known since antiquity. Others were first purified by alchemists such as Hennig Brand (died c. 1692), shown praying after discovering phosphorus.

6

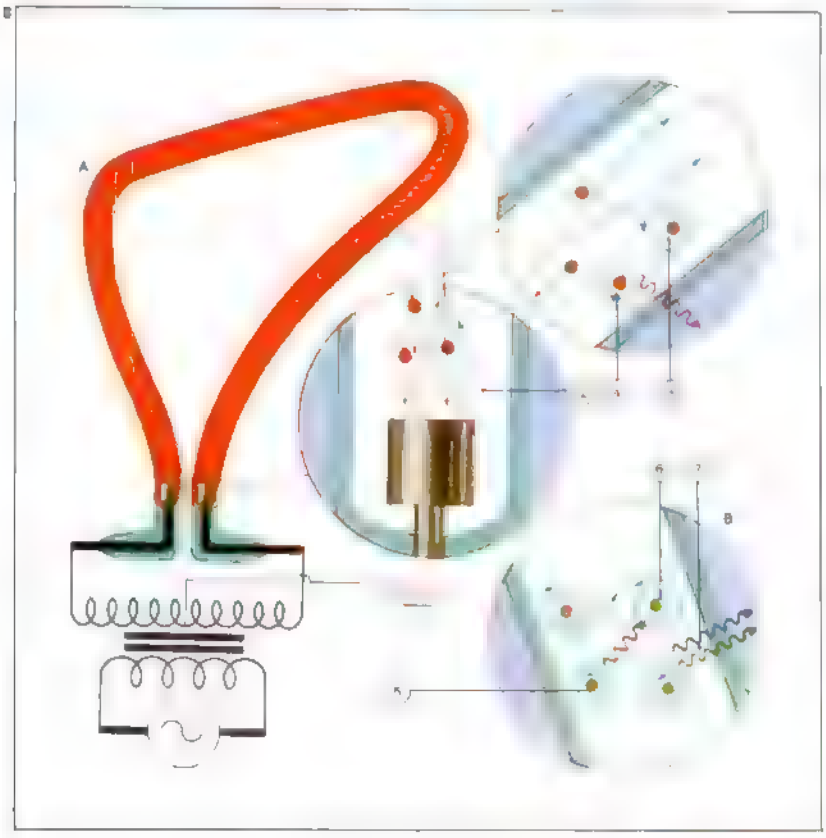
6 Precious metals often occur as the free elements in nature. Not only gold and silver, which for centuries have been used in jewellery and ornamentation, such as this Fabergé egg, but others such as platinum and iridium also come into this category. Silver and mercury are borderline cases. Mercury, for example, occurs as its sulphide but heating releases the metal which readily alloys with other metals.

7



7 A Geiger-Müller counter. [A] detects radioactivity in a chamber [1] containing neon atoms at low pressure [B]. The chamber has neon atoms [grey] that are ionized by beta particles [black] to form positive ions [red] and electrons [blue]. [C] An electric field round an anode [2] accelerates electrons, which collide with other neon atoms [D] and split them to release more electrons, to be recorded by the anode.

8



8 Neon is an unreactive element yet it can be used to produce coloured lights. [A] Positively charged ions [1] strike a negative electrode, causing emission of electrons [2] which "energize" electrons in neon atoms [3]. When the energized electrons return to their stable "ground" state [4], red light is emitted. Similarly, a mercury vapour lamp [B] emits ultraviolet light [5]. This may be absorbed by a fluorescent substance [6] which can then release the energy in steps [7].

9

9 Marie Curie (1867–1934) was the first person to win two Nobel prizes, one in physics, the other in chemistry. She discovered polonium and radium and was responsible for much of the early research on radioactive elements.

Joining atoms

Human beings and the world in which they live exist because atoms of elements join together to form compounds or molecules [Key]. Such compounds may have as few as two atoms or they may contain thousands linked together. From the basic building blocks – consisting of fewer than 100 elements – natural processes have produced hundreds of thousands of compounds and chemists have synthesized many more

Every atom consists of a small central nucleus surrounded by a "cloud" containing one or more electrons. When two atoms approach closely their clouds of electrons interact. Each electron has a negative electric charge and for this reason the electron clouds repel one another. But when two atoms are close together the electrons of each, located in space in "orbitals", are attracted by both the atomic nuclei

Forming a bond

The net result of these dual forces of repulsion and attraction can be a rearrangement of the electron orbitals to form new orbitals that encompass and hold together both nuclei When this happens a chemical bond [1] has been formed and a molecule created. The electrons involved in the chemical bond now occupy molecular, not atomic, orbitals.

The rules that apply to electrons occupying atomic orbitals also apply to those in molecular orbitals [4]. Using the normal "pigeon-hole" notation for showing how electrons occupy atomic orbitals it is also possible to show how electrons are distributed in chemical bonds

In the commonest form of chemical bond, in which a single "pigeon-hole" is involved, the molecular orbital contains two electrons Such a bond, in which two electrons are shared between two nuclei, is called a single, covalent bond [3]. In some cases an atom may have more than a single electron available for bonding. If it meets up with a similar atom, two (or more) covalent bonds may form between the two nuclei

The number of electrons in an atom that are available for forming chemical bonds depends on its outer electronic structure. (All the inner electrons are in completely filled atomic orbitals, which generally cannot accept any more.) For this reason elements such as neon and helium that have completely filled outer shells are highly unreactive and form practically no compounds

Complete electron shells are very stable structures, and there is a tendency for atoms to borrow or share as many electrons from other atoms as they need to complete a shell If two oxygen atoms meet, for example, to fill their outer shells they each take a share of two electrons from the other As a result, each nucleus has eight (instead of six) electrons

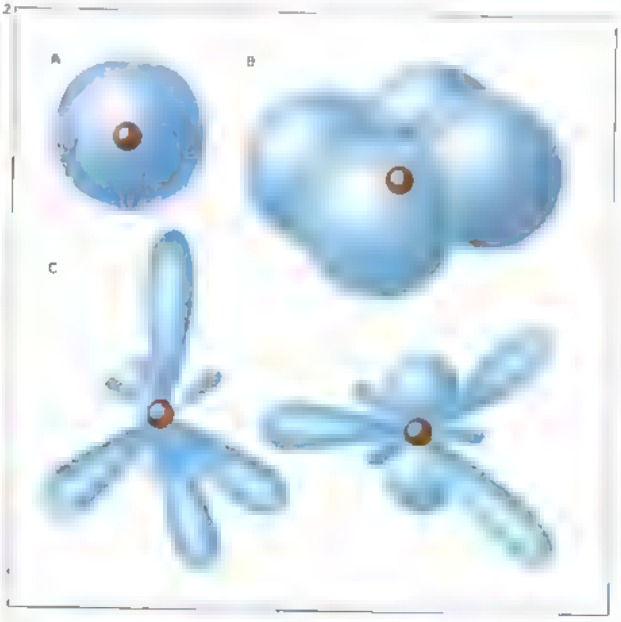
Hydrogen needs – and can accept – one electron to complete its shell (it has only one) [4]. This means that if oxygen and hydrogen atoms come together, two hydrogen atoms each form a single bond to oxygen. In the resultant compound, H_2O (water), the oxygen shares two of its own electrons and has a share in two others, so it again completes its outer octet in the molecule

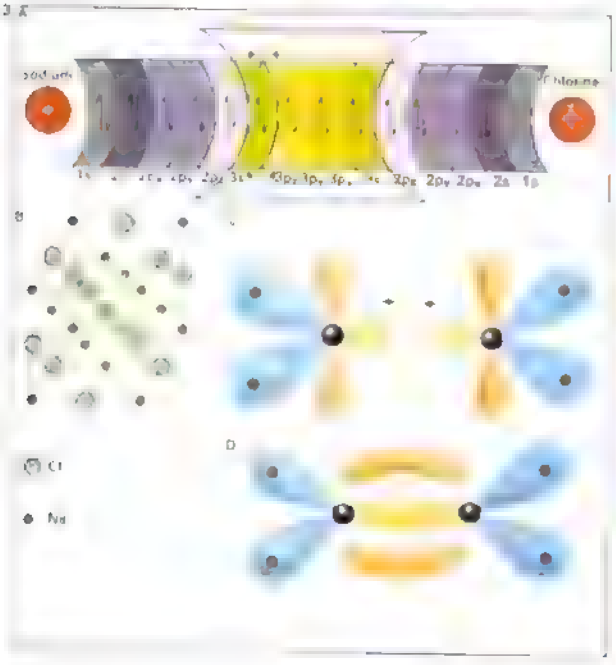
Exchanging electrons

When two electrons are shared between atoms the bond is said to be covalent [4, 6] But some atoms have a stronger affinity for

CONNECTIONS

See also
What is an atom
Simple chemicals and their structures
Complex chemicals and their structures

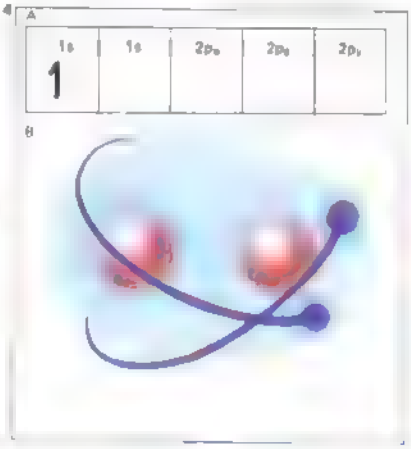




1 When a chemical bond forms the result is often a more stable configuration of atoms. In reaching such an arrangement energy may be released Hydrogen and oxygen can join together to form water molecules, liberating an enormous quantity of energy as they do so This is why hydrogen is so inflammable and no longer used in airships, as it was in the *Hindenburg* which exploded in 1937.

2 The nucleus of any atom has its complement of electrons distributed in orbitals, or volumes of probability For hydrogen's single electron the orbital is spherical [A] while the two available electrons of oxygen are each distributed in a pair of spherical probability regions [B]. The outer electrons of carbon can adopt a variety of distributions such as the two shown here (C).

3 In an ionic bond [A] electrons are transferred, eg from the 3s orbital of sodium (Na) to the 3p orbital of chlorine (Cl) Ionic compounds often have distinct geometries in which ionic charges balance [B] But in a covalent bond two electrons in a molecular orbital are shared by both nuclei Such orbitals have specific shapes as in the 'double bond' (four electrons) linking the carbon atoms in ethylene (C, D) The other carbon orbitals are bound to hydrogen



4 The simplest molecule, H_2, is made up of two covalently bonded hydrogen atoms. The two nuclei share each other's single 1s electrons [A] As two hydrogen nuclei come together [B], and electron sharing commences, energy is released. The positively charged nuclei repel one another and partly counteract the binding force of the electrons This keeps the nuclei a roughly constant distance apart – called the bond length Such two-atom molecules are called diatomic

electrons than others and completely gain or lose electrons to combine by means of an ionic bond. The resulting atoms are no longer electrically neutral, they carry a positive or negative electric charge (depending on whether they have lost or gained electrons) and are known as ions. Chlorine, for example, with seven electrons in its outer shell needs only one more to achieve the stability of a completely filled shell (with eight electrons). When it gains this electron it becomes a chloride ion (Cl^-) and carries a negative charge. Sodium, by contrast, has only one electron in its outer shell and as a result readily loses this to form a positively charged sodium ion (Na^+)

Sodium chloride, or common salt, is an example of a compound with an ionic bond When it is dissolved in water the chloride ions each carry an extra electron borrowed completely – not shared – from the sodium ions

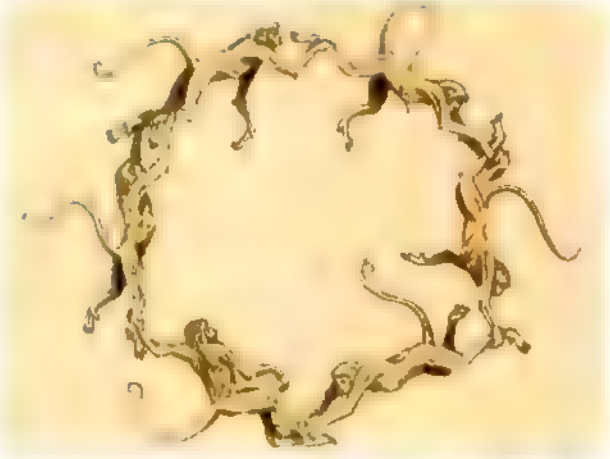
The co-ordinate bond

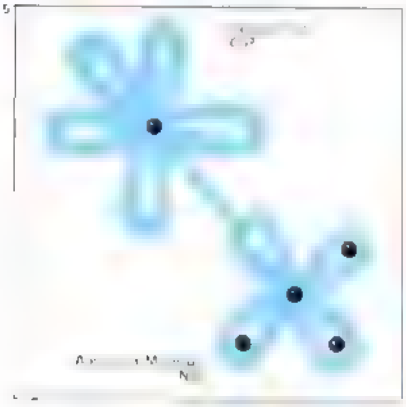
A third possibility exists for joining atoms in a simple way. This is an extension of the sharing arrangement of covalent bonding But in this case both electrons in the bond, called a co-ordinate bond, come from a single atom and use a totally empty orbital in the second atom involved in the bond Nitrogen, for example, has five electrons in its outermost shell. Three of these may be involved in covalent bonds with three hydrogen atoms, to form the compound ammonia, NH_3 Nitrogen now has shares in eight electrons a complete outer octet. In a molecule of ammonia, however, there remains a pair of electrons in the nitrogen atom that have not formed covalent bonds with hydrogen atoms These can be donated to an empty orbital of various metal atoms, such as copper, to form a third type of bond called a co-ordinate bond. The resulting compounds are known as complexes or co-ordination compounds

When more than two atoms are involved in joining together, more complicated arrangements are possible Molecular orbitals may spread out over more than two nuclei to give compounds in which the average number of electrons per bond is less than two, but in which the molecules nevertheless hold together

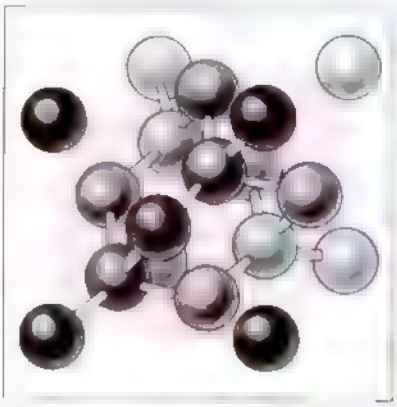
KEY

Atoms join together to form molecules. Only since the discovery of the wave-nature of matter in the 1930s has it been possible to describe this bonding of atoms clearly Before this time the way in which atoms held together was a mystery, often shown fancifully, as in this "model" of benzene Today benzene is pictured as a hexagonal ring of six carbon atoms all linked together by equivalent bonds



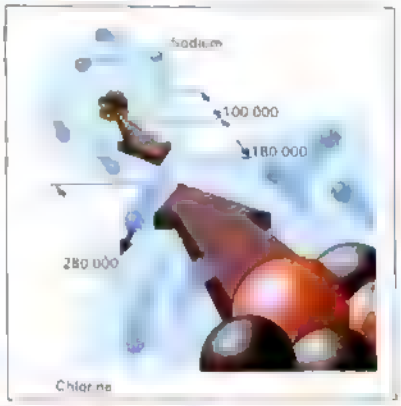


5 A co-ordinate bond involves the sharing of two electrons both contributed by the same combining atom or molecule (hence the alternative name dative covalency) Here a "lone pair" of electrons in an orbital of ammonia enter an empty orbital of a copper (cupric) ion to form such a bond with the required number of two electrons. The resulting compounds are called co-ordination compounds and often are brightly coloured



6 In a diamond any one carbon atom is covalently bonded to four others, orientated at the corners of a tetrahedron. This strain-free covalent configuration accounts for the diamond's hardness.

7 By mimicking the strain-free structure of diamonds chemists have been able to "engineer" even tougher compounds, such as the tungsten carbide on the tips of drills used in oil wells.



8 A sodium atom here appears as it would look seen from the nearest chlorine atom in a crystal of sodium chloride One electron is shown moving from the sodium to the chlorine, resulting in an ionic bond between them The dimensions of the sodium nucleus (made up of protons [red] and neutrons) and of the two complete atoms are given in femtometres – a unit equal to a million millionth of a millimetre (10^{-15}m)

9 When two atoms spontaneously form a bond the release of energy stabilizes the compound The process may be reversed by supplying energy. Colourless silver salts used in automatically dimming sunglasses (A) and an astronaut's helmet (B) take the form of silver ions in a glass matrix. In sunlight the glass darkens because light energy breaks the ionic bonds and reconverts the silver into metallic atoms. Colourless ions re-form when the light is removed

10 Chemical bonding can be illustrated in many ways. The four diagrams shown here depict the simple molecule ethylene, made up from two carbon atoms and four hydrogen atoms As in the diamond structure, each carbon atom can form four bonds two of these link the two carbons together and each carbon also has two hydrogens attached to it by single covalent bonds The simple ball-and-stick model A shows the atoms and the number of bonds linking them. Model B gives similar information but is designed specially for biological molecules while models C and D reveal more about the actual molecular shape, and the distribution of orbitals, respectively

Simple chemicals and their structures

Mankind's interest in chemistry derives from the useful information that the subject can provide about the properties of different substances. By understanding the structure of molecules and the way they interact with one another it is possible to invent new compounds that are useful, for example, as drugs or building materials or fibres for clothing. It is also possible to acquire a better understanding of how the earth on which we live came to its present form and how it is still developing. Geochemistry (as this sub branch of the subject is called) can lead to the discovery of new supplies of fuels and ores and new ways of processing them

Structures of molecules

All substances are made up of molecules that in turn, are composed of individual atoms One of the basic aspects of chemical knowledge is an understanding of the structure of molecules. An atom is a relatively insubstantial entity: a small, hard nucleus surrounded by an electron occupying a volume of probability Molecules composed of such atoms are similarly made up, in terms of volume, largely of electron orbitals (the areas in space occupied by the electrons). Nevertheless despite the fact that an orbital can contain a maximum of only two electrons, they often have definite directions in space, so that molecules have particular shapes. In complex molecules such shapes may be crucial to the behaviour of the substance This is particularly true of biological molecules, which contain thousands of individual atoms linked together but even simple molecules have shapes that can determine their properties

The structures of ionic compounds [1] depend on the electric charge of the ion and on its ionic volume An ion of sodium, for example, can be regarded as a small sphere having a particular diameter. In general, provided ions of different elements have the same electric charge and similar radii, they can substitute for one another in different materials As the earth settled down after its formation, rocks gradually solidified from molten material Many of these original rocks contain crystalline compounds but the compounds are nearly always impure. For example, the common mineral potassium felspar, an ionic compound containing potassium, silicon and oxygen, is always contaminated with rubidium because rubidium forms a singly charged ion of approximately the same radius as potassium Germanium forms a similarly charged, similarly sized ion to silicon so it is also a contaminant of felspars

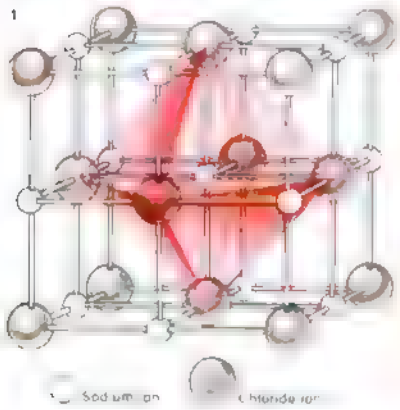
In covalent (non-ionic) compounds the bonds between the different atoms are separate Covalent compounds can still form crystals, but this is because the molecules as a whole can arrange themselves in geometric arrays, whereas with ionic compounds it is the geometric arrangement of ions that leads to crystallinity With a crystal of a covalently bonded compound each component molecule has its own characteristic shape

Shapes of covalent compounds

The shape of a covalently bonded molecule [3] depends on the shapes of the orbitals occupied by the electrons – both those involved in chemical bonding and any others in the outermost shells of the individual atoms A water molecule, for example, in which two hydrogen atoms are each singly

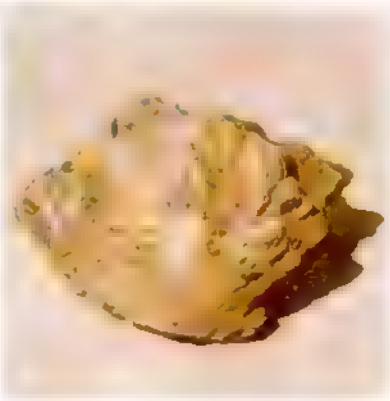
CONNECTIONS

1



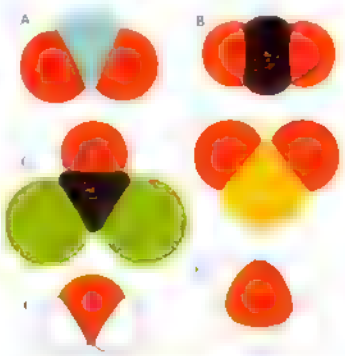
1 Sodium chloride, common salt has a simple cubic crystal structure in which each sodium ion (Na^+) is surrounded by six chloride ions (Cl^-) and vice versa Around each ion, the six ions of opposite electrical charge are situated at the corners of an imaginary octahedron The main force holding the ions in place is the balanced electrical attraction of the neighbouring ions

2



2 The shape of any crystalline, ionic solid depends on the size and the number of ions that make it up However different chemical substances may look alike, even though they have different compositions, molecular structures and chemical properties Iron sulphide shown here, is called "fools' gold" because of its resemblance to the much more valuable element, gold

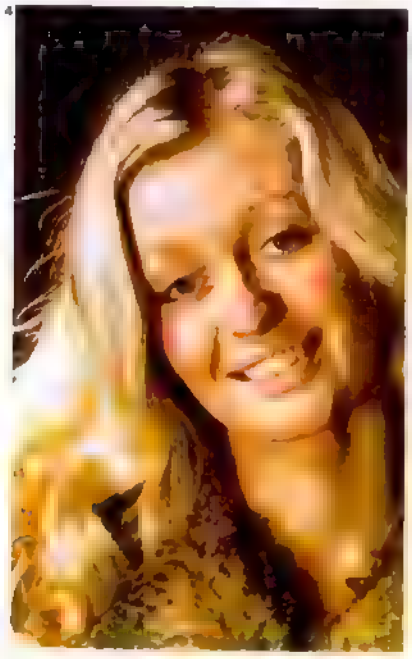
3



A Nitrogen dioxide NO_2
B Carbon dioxide CO_2
C Phosgene $COCl_2$
D Sulphur dioxide SO_2
E Water H_2O
F Ammonia NH_3

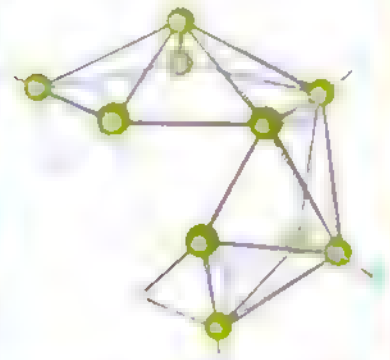
3 Covalently bonded molecules have individual molecular orbitals holding different atoms together unlike the balanced geometric cluster of an ionic compound Because of the forces of repulsion between the electron clouds surrounding adjacent nuclei, even small molecules have distinct shapes, which may differ quite markedly from one another as in the half-dozen common molecules here

4



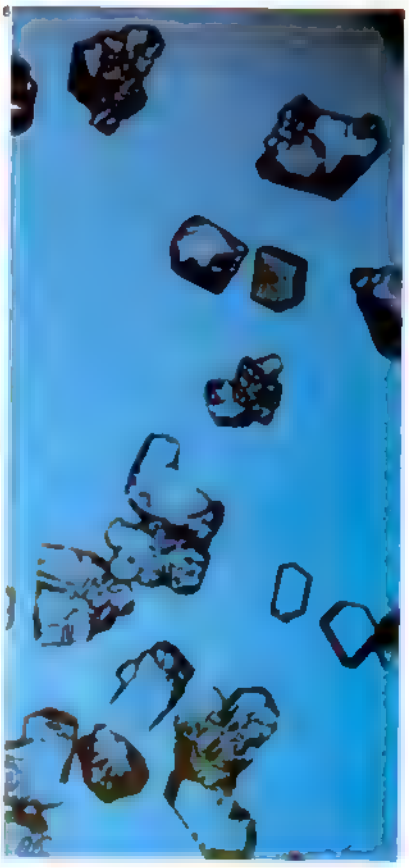
4 In hydrogen peroxide two hydrogen atoms and two oxygen atoms are joined in a single molecule It readily releases an OH group whose oxygen can, for example, bleach hair (as here) or kill germs and act as a mild disinfectant

5



5 In most covalent bonds two electrons hold two nuclei together By adopting tightly geometric patterns electron deficient compounds such as this decaborane can hold together with a smaller number of electrons

6



6 In addition to electron-deficient compounds, there are also compounds that seem to have too many electrons. For many years it was believed that the noble gases – helium, neon, xenon, krypton and radon – would not form any compounds because their outer electron shells had eight electrons in them already However in 1963 Neil Bartlett discovered that xenon would react to form colourful crystalline compounds. An extension of chemical bonding theory has shown how the formation of such compounds does not violate any chemical law and since 1963 a large number of compounds of xenon (shown here) and krypton have been made in laboratories throughout the world

bonded to a central oxygen atom, might be visualized as the three atoms joined in a straight line. Electrons, all being negatively charged, repel each other, so this in-line arrangement might seem best. It would mean that the volumes of electron probability around the two hydrogen atoms are as far apart from each other as they possibly could be, thus reducing repulsion to a minimum. But, in addition to the two electrons from the oxygen's outer shell, which are involved in bonding, there are four other electrons in this shell, situated as two "lone pairs" in filled orbitals. Their effect on neighbouring electron clouds has to be taken into account. When this is done it is found that a shape almost like that of one segment of a diamond [3E] is adopted. It is a tetrahedral structure in which the hydrogen atoms and the two pairs of oxygen electrons (not involved in bonding) all lie as far away from each other as possible.

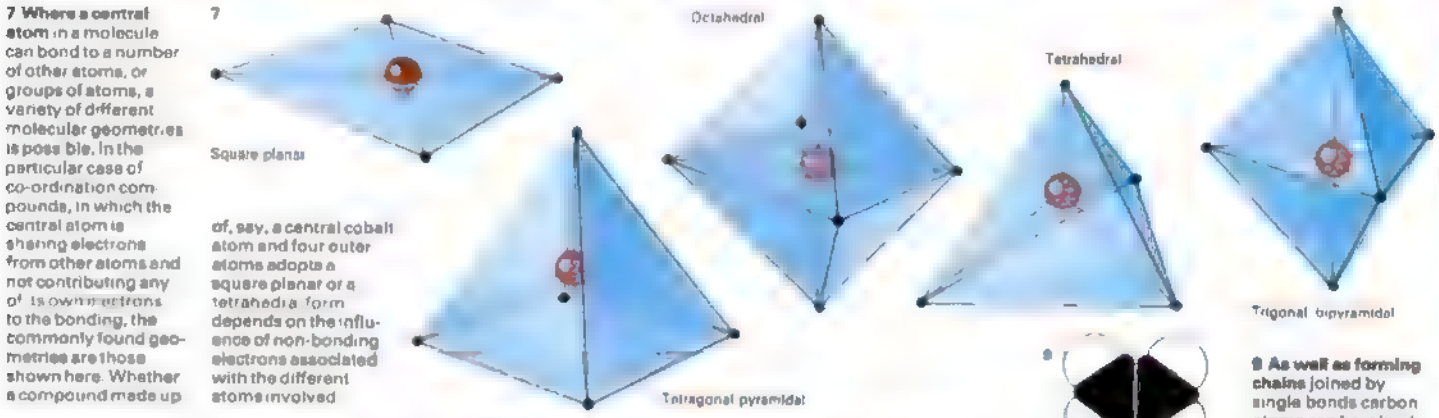
Methane [8], in which one carbon atom is surrounded by four hydrogen atoms and all the outermost carbon electrons are involved in bonding, adopts a regular tetrahedral configuration. Intermediate in structure between methane and water is ammonia (NH_3), in which there are three nitrogen-hydrogen bonds and one filled orbital containing a lone pair of nitrogen electrons [3F].

Co-ordination compounds

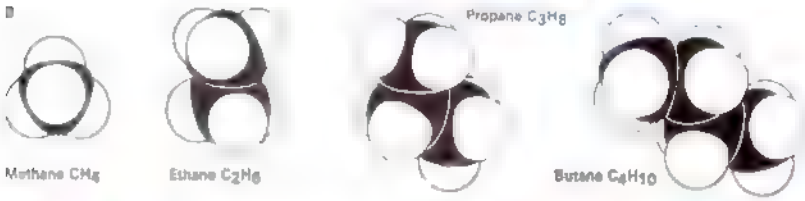
The lone-pair electrons in water and ammonia molecules sometimes form bonds to metal atoms that have empty orbitals. Anhydrous copper sulphate, for example, is white. When water is present the molecule turns blue as a number of water molecules "co-ordinate" to empty copper orbitals through the unbonded oxygen lone pairs. Because the empty orbitals in such metals have definite shapes most "co-ordination compounds" generally have highly geometric structures [7].

Some atoms can form several bonds with other atoms, so it is often possible for a number of different molecules to be made up from the same mixture of atoms [10]. One of the major advances in chemistry since the 1940s has been the development of sophisticated instruments that analyse exactly how the atoms of a compound are linked together.

Everything around us is made up of chemicals, most of them complex mixtures. But these familiar substances from the kitchen are comparatively simple chemical compounds.

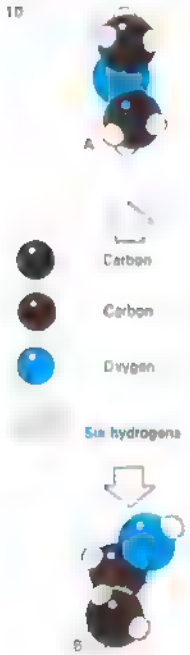


7 Where a central atom in a molecule can bond to a number of other atoms, or groups of atoms, a variety of different molecular geometries is possible. In the particular case of co-ordination compounds, in which the central atom is sharing electrons from other atoms and not contributing any of its own electrons to the bonding, the commonly found geometries are those shown here. Whether a compound made up of, say, a central cobalt atom and four outer atoms adopts a square planar or a tetrahedral form depends on the influence of non-bonding electrons associated with the different atoms involved.



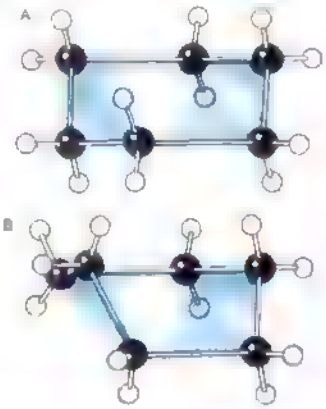
8 Carbon atoms can combine covalently to form long chains. Consequently carbon compounds occur in series such as the alkanes (paraffins); the first four (commencing with methane) are shown here.

9 As well as forming chains joined by single bonds carbon atoms can be joined to one another by two or three molecular orbitals. ethylene and acetylene are, respectively, the doubly and triply bonded analogues of ethane, all of which contain two carbon atoms.



10 Where the same numbers of the same kinds of atoms are joined to one another in different ways the resultant compounds behave differently. Because the gross structures of the molecules are different they are called structural isomers; examples are dimethyl ether [A] and ethyl alcohol [B].

11 Some compounds such as cyclohexane, can exist in two forms differing only in the way their chemical bonds are arranged in space. This illustration shows the "chair" [A] and "boat" [B] forms of cyclohexane. Such compounds are known as conformations: they may "flip" from one form to the other, and a sample of the compound may contain both.



12 During the 20th century hundreds of thousands of carbon compounds hitherto unknown in nature have been synthesized in chemical laboratories. Some of the simplest such molecules of commercial significance are insecticides. Developed mainly to protect food crops, they are also a help to the home gardener. The active ingredients in many insecticide sprays and powders – dichlorvos and BHC for example – are molecules containing only one or two dozen atoms.

Complex chemicals and their structures

Carbon atoms have the ability to link together in large numbers to give an infinite variety of different substances [2]. Most complex chemicals are carbon-based although there are some important complex materials, such as glass, that contain no carbon.

At one time it was believed that most carbon compounds could be made only by living processes. For instance urea was discovered in the urine of mammals. Consequently, this and similar compounds were called organic chemicals, a name that has stuck. Then organic chemicals were synthesized in the laboratory – for example urea (organic) was made from ammonium cyanate (an inorganic compound). Today organic chemicals account for 50 per cent of the total output of the chemical industry.

The basis of an organic chemicals industry

The discoveries that led to the development of an organic chemicals industry came towards the end of the nineteenth century [3]. It was found that a wealth of useful substances could be obtained from coal tar, a by-product of the manufacture of domestic gas from coal. At that time many of the substances extracted from coal tar were too complex in structure to be made in the laboratory. Once purified, however, it was possible to use them as starting materials from which to produce a range of commercial substances. Dyes, aspirin, saccharin, and explosives such as TNT were all made before the end of the nineteenth century from coal-tar chemicals.

A basic constituent of many of the coal tar products, and of many other complex organic chemicals, is the group formed of a ring of six carbon atoms joined together. As in other molecules, the chemical bonds between the carbon atoms take the form of electron-carrying orbitals. These orbitals are not located strictly between adjacent carbon nuclei but spread over all six, so that each carbon atom is effectively joined to the next by one-and-a-half bonds.

The simplest of such compounds, in which each carbon atom is also linked to a hydrogen atom (giving a formula C_6H_6), is benzene, originally a coal-tar product but now made mainly from petroleum [1]. In some compounds several benzene rings are fused together to give what are called polycyclic structures such as naphthalene (used in mothballs), benzpyrene (a cancer-inducing chemical) and the hallucinogenic drug LSD.

Many complex organic chemicals found in nature do not contain benzene-type rings. Instead, they are made up of long chains of carbon atoms, with other atoms attached. The other atoms nearly always include hydrogen and often oxygen and nitrogen as well. These compounds include useful natural products such as fats, waxes, sugars and proteins.

Synthetic polymers and their products

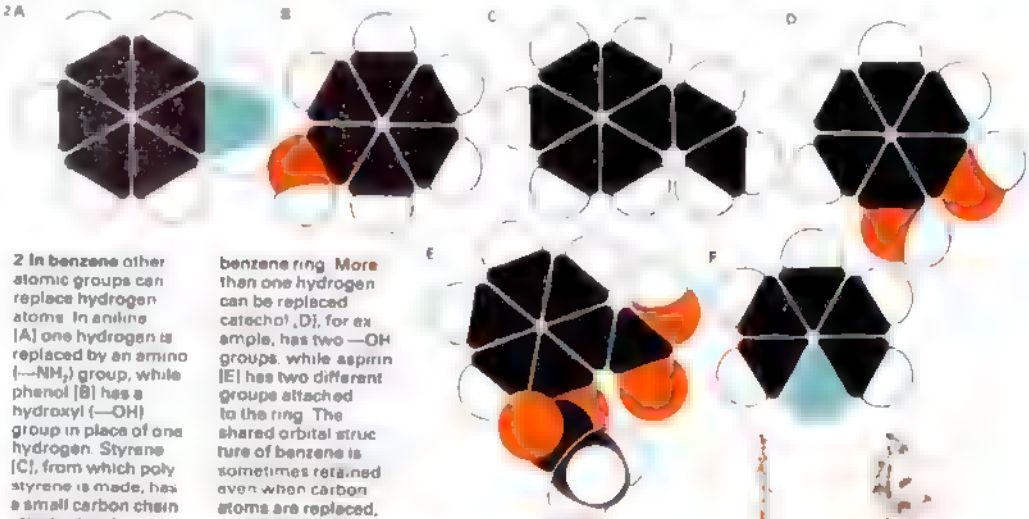
Organic molecules have been the basis of one of the major industrial developments of the twentieth century – the widespread manufacture and use of synthetic polymers as plastics, rubbers and fibres. Polymer is a general term for any large molecule that is made by repeatedly linking together the same small molecular unit, which is called the monomer.

Most organic polymers are formed either by addition reactions or by elimination reactions. In many examples of the first type the monomer molecule has a double bond and

CONNECTIONS

See also

1 The six-carbon benzene ring is one of the most important molecules in organic chemistry. Each carbon atom is attached to a single hydrogen atom, so it can form three other bonds with its two neighbouring carbon atoms. It was once thought that single and double bonds alternated around the benzene ring [A, B], although if this were so not all the carbon atoms would have the same chemical reactivity. It is now known that all the carbon-carbon bonds in the ring are equal because the molecular orbitals (blue [C]) spread out over the whole ring. Below [D] is a more convenient representation of the actual shape of the molecule.



2 In benzene other atomic groups can replace hydrogen atoms. In aniline [A] one hydrogen is replaced by an amino (—NH_2) group, while phenol [B] has a hydroxyl (—OH) group in place of one hydrogen. Styrene [C], from which polystyrene is made, has a small carbon chain attached to the benzene ring. More than one hydrogen can be replaced: catechol [D], for example, has two —OH groups, while aspirin [E] has two different groups attached to the ring. The shared orbital structure of benzene is sometimes retained even when carbon atoms are replaced, as in pyridine [F].

3 The English chemist William Perkin (1838–1907) tried to make quinine from aniline. Instead his experiments of 1856 accidentally produced the dye aniline purple. This was the first synthetic dye: until that time all dyes were natural compounds extracted from plants or animals.

4 Modern pigments brighten this train with contrasting colours, and paint makers are no longer dependent on natural substances for their products.

5 The chemical heart stimulant digitalin was first obtained by herbalists from leaves of the foxglove (*Digitalis purpurea*).

links with others by using electrons from the double bond. In the second type atoms at the ends of two monomer molecules are "eliminated" and a bond forms between the remaining parts of the monomer molecules.

Natural rubber appears to be an addition polymer, as are many of the man-made rubbers developed in the past few decades. Many *synthetic fibres on the other hand are* elimination polymers or condensation polymers (in which water is eliminated), as are their natural counterparts such as wool and cotton. A synthetic polymer can have characteristics built into it by the careful choice of starting materials. As a result, different synthetic fibres can have widely differing properties.

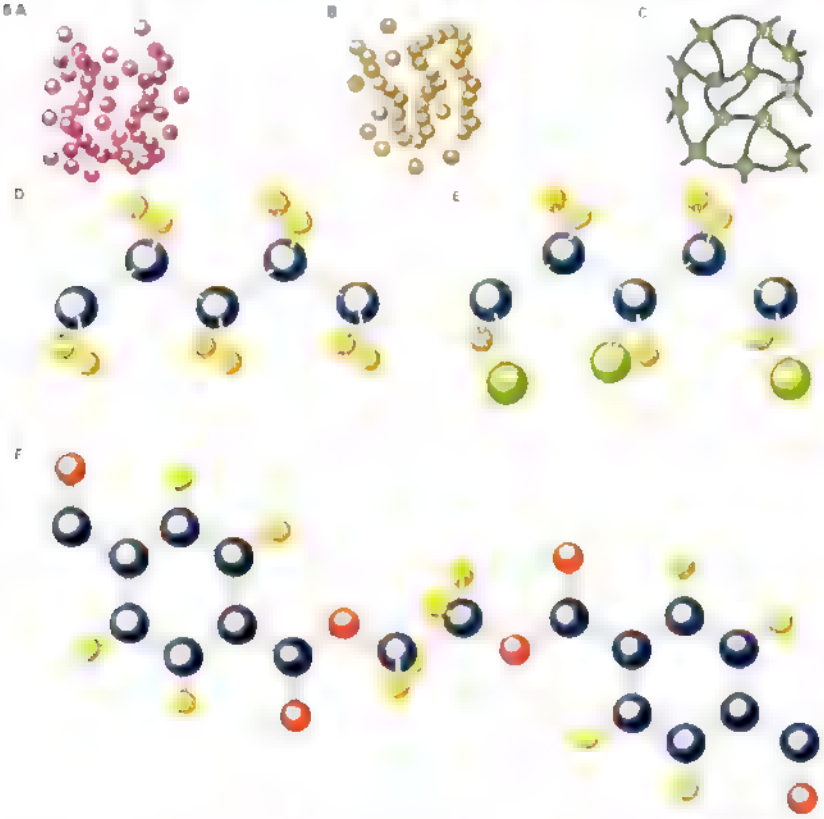
Plastics [6] can be fibrous, rubbery, clear, hard, opaque or flexible. The possibilities are almost endless, as the widespread use of synthetic polymers in everyday life confirms.

Linking silicon atoms

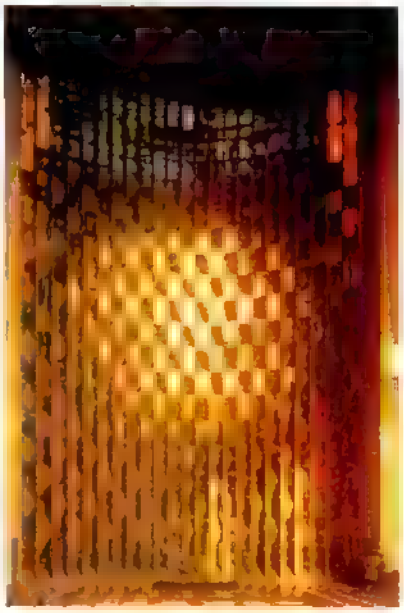
Silicon falls directly below carbon in the periodic table of the elements and so logically should have similar chemical properties. But because of their larger size silicon atoms cannot link to form long chains by themselves. Nevertheless, the important polymers called silicones are based on long chains of alternating atoms of silicon and oxygen. The commercially available silicones are often partly organic, and their forms range from plastics for replacement parts such as artifi*cial heart valves in surgery, through greases* and lubricating fluids, to insulating materials for electrical cables in submarines.

In silicones, carbon groups are linked to the silicon atoms and stick out on each side of the polymer chain. Like carbon, silicon can form four bonds, whereas oxygen forms only two. This means that, in a repeating —Si—O— chain, there must be other atoms joined to the silicon atoms. These do not have to be organic: each silicon atom can be bonded to four oxygen atoms. If each of these oxygens has two silicons attached, the result is a three-dimensional polymer matrix with the overall formula SiO_2. Sand and glass are composed largely of this inorganic polymer [7], which is often known as silica, and it is also used in ceramics and paint manufacture.

A modern chemist deals in hundreds of drugs. The old chemist, called an apothecary, displayed in his shop rows of glass jars filled with mixtures and essences extracted from botanical plants. *Many balms and* medicines made from the contents of such jars are still manufactured today, not from natural extracts but from substances that are produced in chemical plants. Increasing knowledge of the complex molecules found in nature has enabled man to copy and in some cases improve nature's own efforts. Synthetic drugs, rubber and pigments are only a few of the products that man's ingenuity has created to replace natural materials.

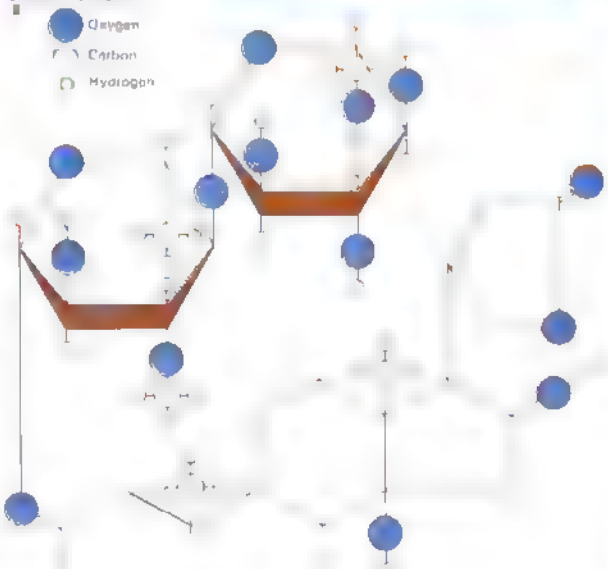


6 Plastics fall into three main structural groups: polymers, copolymers and thermosets. A polymer [A] consists of repeated units of the monomer built into a chain. A copolymer [B] has units of more than one monomer. Thermosets [C] form cross links between the chains. Both polyethylene [D] and PVC [E] are typical polymers. Polyesters (such as Terylene) [F] cross link on heating to give thermosets.



7 Colours in glass, such as those in the windows of Coventry Cathedral, come from added metallic compounds. Cross-linked polymers are not all based on organic molecules and glass is probably the oldest synthetic polymer.

8 Modern plastics are replacing metals in many applications, such as the bodywork of motor vehicles.



9 Digitalin, one of the active ingredients derived from the foxglove (see illustration 5), is typical of biologically active complex compounds found in nature. The three-ring phenanthrene residue in the lower half of the molecule is common also to such compounds as steroids and hormones, cholesterol and the human male and female sex hormones, for example.

Chemicals in solution

Every day millions of people make solutions. Many start at breakfast, when they dissolve sugar in tea or coffee. So common is this action that it is taken for granted. But where does the sugar go when it dissolves? Why, if a spoonful of sand is stirred into hot liquid does it not dissolve?

In chemical terms, a solution is a homogeneous mixture of different sorts of molecules. The criterion of homogeneity is that the two or more types of molecules involved are thoroughly mixed up with one another – as water and sand clearly are not.

Solutes and solvents

Solutions are usually thought of as solids dissolved in liquids – coffee in water, sugar in the coffee, salt in water in the sea [6], detergent in washing water, iodine in alcohol in tincture of iodine, and many more. But there are other kinds of solutions. Gases can dissolve in liquids, as in soda water. Many solutions are made by dissolving liquids in other liquids. Gases can also dissolve in some solids, and solutions of solids in solids are found in metal alloys, for example.

For a solution to form, there must be an interaction between the substance which dissolves (solute) and the material it dissolves in (solvent). Sugar, for example, usually occurs as crystalline arrays of sucrose molecules. To dissolve it, energy is needed to break apart the crystal lattice so that sucrose molecules can disperse evenly through the solvent. Where water is the solvent, the attraction for solute molecules comes from the "polar" character of the water molecule. The central oxygen atom in H_2O is electrically slightly negative, while the hydrogen atoms are slightly positive. Molecules of water tend to attract one another – which is why water is a liquid at room temperature whereas most such small molecules are gases [1]. Water molecules also tend to attract (ie dissolve) other "polar" compounds, such as sugar, with many OH groups in their molecules.

Molecules and compounds

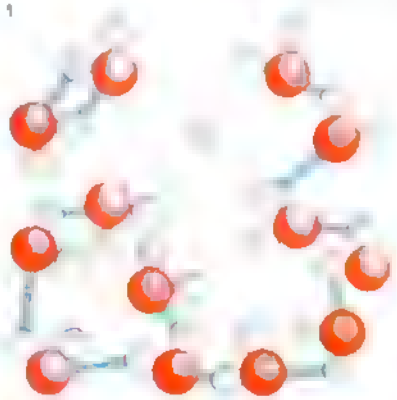
For similar reasons, but via weaker attractive forces, non-polar molecules such as hydrocarbons will dissolve other non-polar compounds, such as fats. Modern detergents work by a compromise between both types of attraction: part of the molecule dissolves in grease, the other part dissolves in water, so that the detergent molecule acts as a bridge between them and disperses grease in water.

Some compounds, such as ethyl alcohol, dissolve completely in either water or hydrocarbons – substances that will not dissolve in one another. Other compounds may show a preference for polar or non-polar solvents, according to their chemical structures.

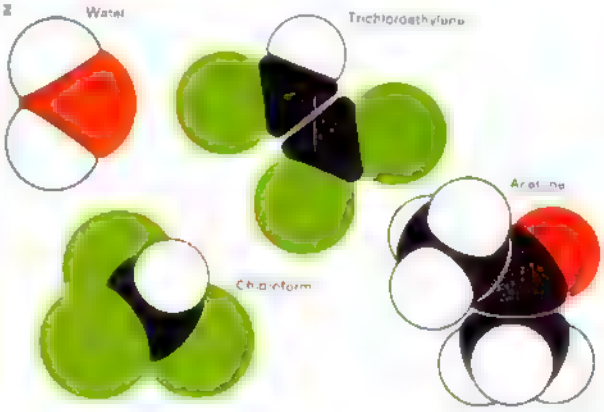
The attraction between individual molecules of different types can be seen easily where it results in a reduction of the overall volume of material. If, for example, equal parts of water and ethyl alcohol are mixed together, the total volume of solution is only about 97 per cent of the sum of the separate components.

When a substance will not dissolve, it is because the solvent does not overcome the intermolecular forces which hold the molecules together. With any substance there are limits to solubility, which may vary with temperature – hot solvents dissolving more solute than cold solvents.

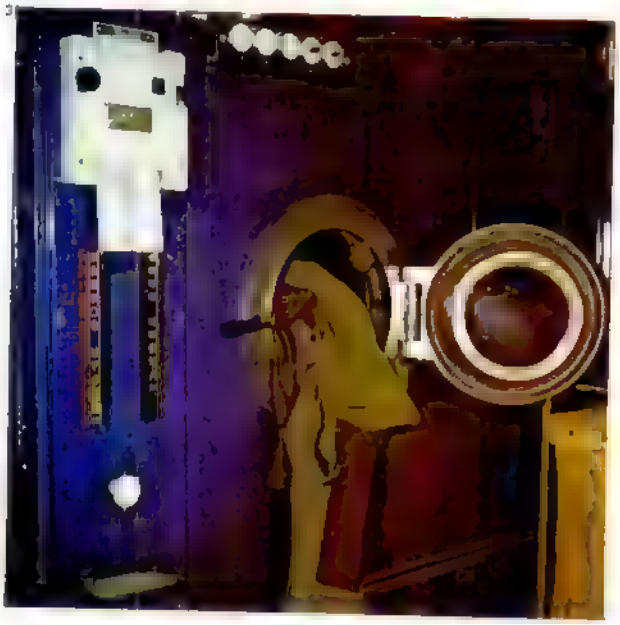
CONNECTIONS



1 The atoms and molecules of a liquid are in motion: the energy of this motion depending on the temperature. In a liquid, the motion prevents any permanent intermolecular structure from forming, but forces of attraction govern the overall volume. In water, there are many temporary linkages (shown blue) between the H_2O molecules; very small enclosed cavities [1] form and disappear continually.

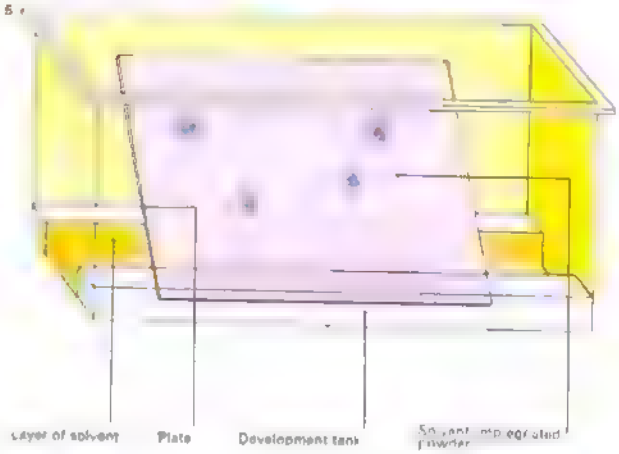


2 The slight positive charge on the hydrogen atoms and slight negative charge on the oxygen atoms of water make it particularly suitable for dissolving inorganic salts. Trichloroethylene, widely used in dry cleaning, and chloroform both dissolve many organic compounds. Acetone, with behaviour intermediate between water and trichloroethylene, is able to dissolve both organic and inorganic compounds.



3 Ordinary washing removes water-soluble dirt and also non-soluble dirt that can be emulsified in water by the addition of detergents. (Milk is a typical example of an emulsion where one liquid mixed and joined with another does not separate.) In dry cleaning, an organic solvent is employed, most of those used commercially are small, halogen containing molecules, such as trichloroethylene. These solvents dissolve grease, however, because their fumes are unpleasant they have to be used in special machines, as shown here. Similar compounds are also used in industrial degreasing operations, such as cleaning pieces of metal prior to machining, or in the manufacture of electronics components.

4 The solubility of a substance is different in different solvents. When immiscible solvents are shaken together, any compounds present are divided between the two liquid phases. If a solution of iodine and salt in water is shaken with benzene, iodine dissolves in the benzene layer but salt does not. Thus it is possible to separate the two. Many different compounds can be purified in this way.



5 Different compounds can also be separated by partitioning them between a solid and a liquid. In thin-layer chromatography, the process shown in this illustration, a spot of mixed compounds is placed on a plate coated with absorbent powder, which is put into a development tank containing a solvent (ensuring the tank is saturated with its vapour). As solvent ascends the plate, the mixture separates out into its component chemicals.

A solution containing the maximum possible amount of solute (dissolved substance) is called saturated. A hot liquid generally has a greater capacity for holding solid in solution than a cold one (with gases in liquids, the reverse is true). Consequently, if as much solid as possible is dissolved in boiling water some of it may crystallize out when the water cools. If it does not, the solution is said to be supersaturated. This is the basis of many experiments for growing giant crystals in school laboratories.

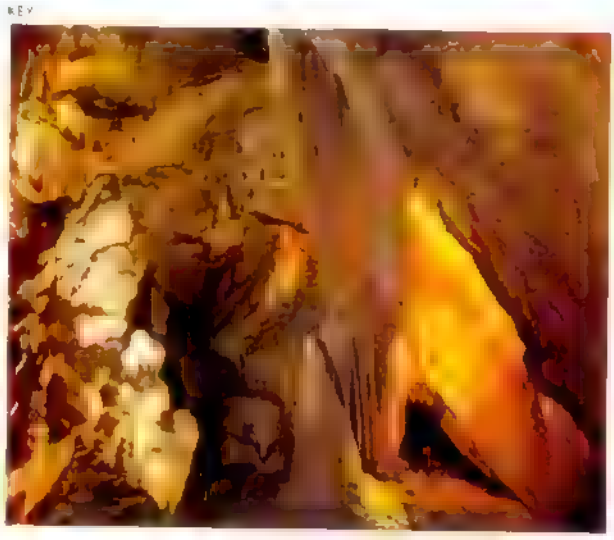
Dissolving a solid in a liquid affects the liquid in a variety of ways. For example, pure water freezes at 0°C (32°F) and boils at 100°C (212°F). The freezing-point of a solution of common salt, however, is lower than 0°C; thus, salt sprinkled on roads in winter prevents ice from forming.

Osmotic pressures

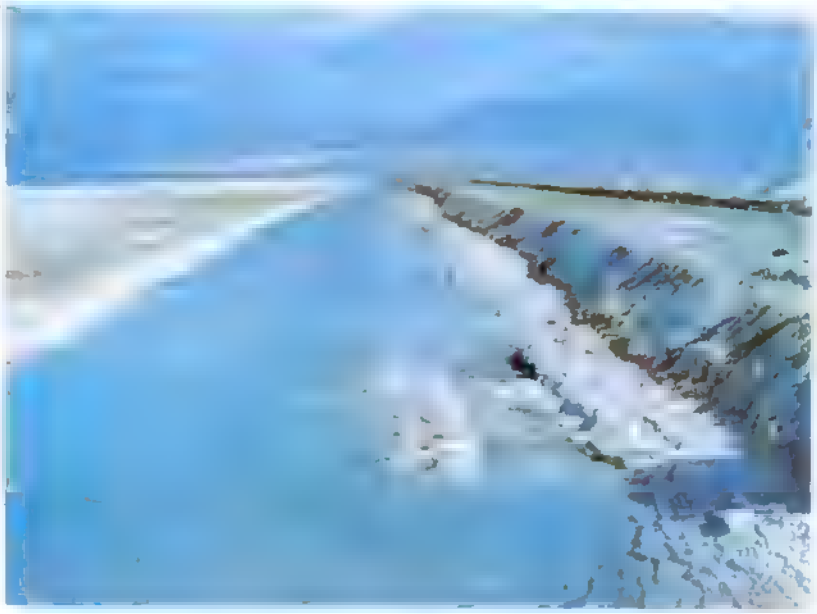
One of the more important properties of solutions is their ability to exert osmotic pressure. If a solution enclosed in certain types of membrane is brought into contact with pure solvent, the solvent molecules pass through the membrane into the solution, making it more dilute. The molecules of solute, on the other hand, are unable to pass through the membrane, which is called "semi-permeable" in consequence.

Osmosis is crucial to many living organisms. For example, the absorption of water by the root hairs of plants depends on it. If the concentration of dissolved matter in the plant cells is greater than in the water surrounding them, water is drawn in. On the other hand, if the reverse is true – in very salty soils, for example – water may be drawn out of the plants, so that they die.

Plants and animals are also responsible for producing complex substances that form so-called colloidal solutions. These are part way between a solution and a suspension. The jellies eaten by children are one type of colloidal solution, and non-drip paint is another. Jellies set on cooling and "melt" again on heating, but a non-drip paint becomes more liquid as it is stirred and more solid when it is left standing. These differences are accounted for by the different types of colloidal substances used.

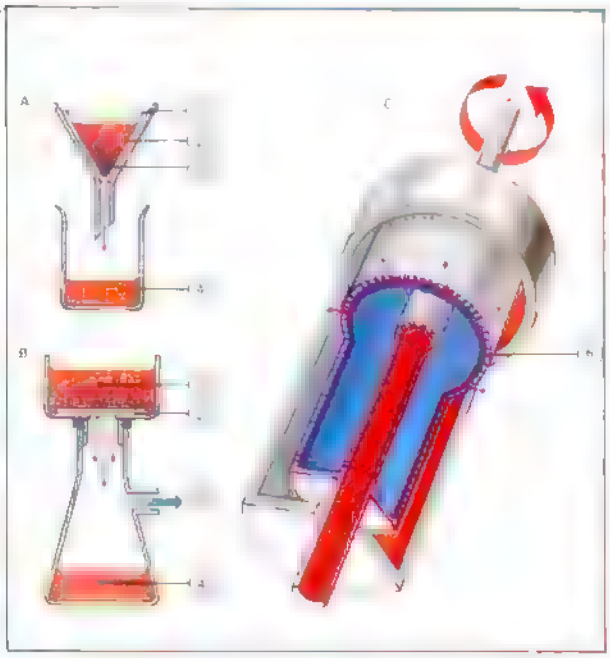


Rivers and streams dissolve small quantities of minerals. Under certain conditions, these minerals can be precipitated from solution – sometimes in impressive forms, such as these stalactites and stalagmites formed over thousands of years in limestone caves in various parts of the world.

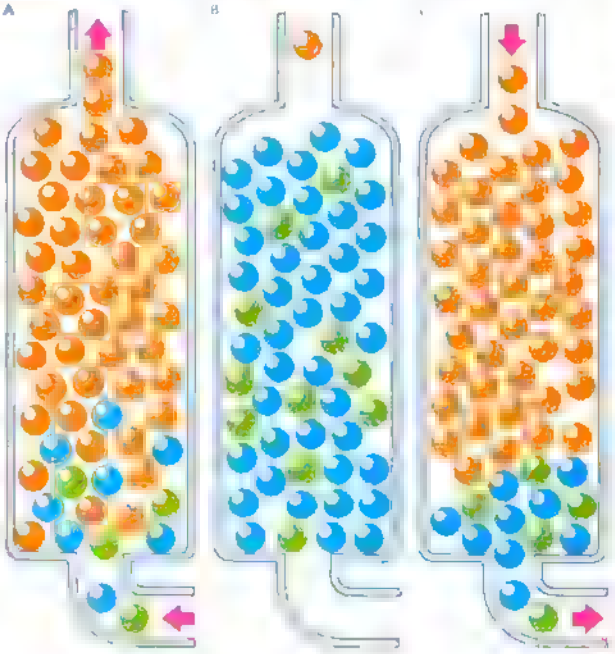


6 There is a limit to the amount of any particular compound that will dissolve in a solvent. When this limit is reached, the solution is said to be saturated. If solvent is evaporated or if the temperature falls the amount of compound that can be held in solution drops and solid precipitates. The Dead Sea is one of the most concentrated naturally occurring solutions of minerals, some of which crystallized out when the water level dropped in the past. Here a chemical industry evaporates large quantities of Dead Sea water in nearby "pans" to obtain important minerals for making various inorganic chemicals.

7 Although two-thirds of the earth's surface is covered with water, in many countries there is a lack of pure water for drinking. In this modern desalination plant, pure water is evaporated from solutions such as seawater and then condensed in large tanks. The concentrated brine left over from the process is usually returned to the sea, where it is diluted by tidal mixing.



8 Filtration removes solid particles from solution. A mixture placed (A) in a filter funnel (1) drains through the filter paper (2) by gravity, leaving sediment (3) trapped while the filtrate (4) flows through. In the Buchner funnel (B), the same principle is enhanced by suction (5). In the industrial rotary filter (C), the spinning action of the drum drives the mixture onto the fine mesh filter (6).



9 In a domestic water softener, water flows up a column (A) and ions of calcium and magnesium (which cause "hardness") are exchanged for sodium ions. In time all the sodium ions are used (B). The column is regenerated using sodium chloride (C).

Key chemical reactions

In an enclosed space, a mixture of air and an organic vapour composed mainly of hydrocarbons are pressed together. The hydrocarbons consist of molecules in which several carbon atoms are chemically bonded to each other and to a number of hydrogen atoms. A spark is generated. In an instant, many of the bonds between carbon and hydrogen and between carbon and carbon are broken and replaced by chemical bonds that combine these atoms with oxygen from the air. An explosion results, and a chemical reaction has taken place. This particular reaction takes place millions of times each day in most parts of the world, wherever people use petrol-driven internal combustion engines in cars and other motor vehicles.

Molecular energy

Chemists are interested in how individual compounds react: how quickly they will react, what the products are and how much "persuasion" is needed to make a reaction take place.

Why, for example, is a spark needed before the hydrocarbons that compose petrol will react with oxygen? Why, however many sparks are provided, will exhaust gases such as carbon dioxide and steam not burn? Why does the spark cause an explosive reaction in a petrol/air mixture, whereas a spark landing on this page would probably only char a small portion of the paper; and, if the paper did catch fire, why would it burn steadily rather than explode?

The answers to all these questions are related to the energies of different molecules. The world is full of molecules, rather than unlinked atoms, because the formation of chemical bonds releases energy and makes the resultant product more stable. The same is true of molecular reactions. If a reaction occurs spontaneously, there is usually a release of energy and the production of more stable molecules.

It is theoretically possible to get water and carbon dioxide to react to form petrol and oxygen. But, because petrol and oxygen are less stable molecules, large quantities of energy would have to be put into the reactants (starting materials) to succeed. A better example is, perhaps, metallic corrosion. Many metals are more stable as compounds, such as the oxide or sulphide, than as pure metal. Using processes that supply energy to metallic compounds, scientists can refine their ores to make metals for steel girders or silver teaspoons. But if unprotected, the metals gradually corrode – that is, spontaneously form compounds such as iron oxide (rust) or silver sulphide (tarnish) that are more stable than the pure metals.

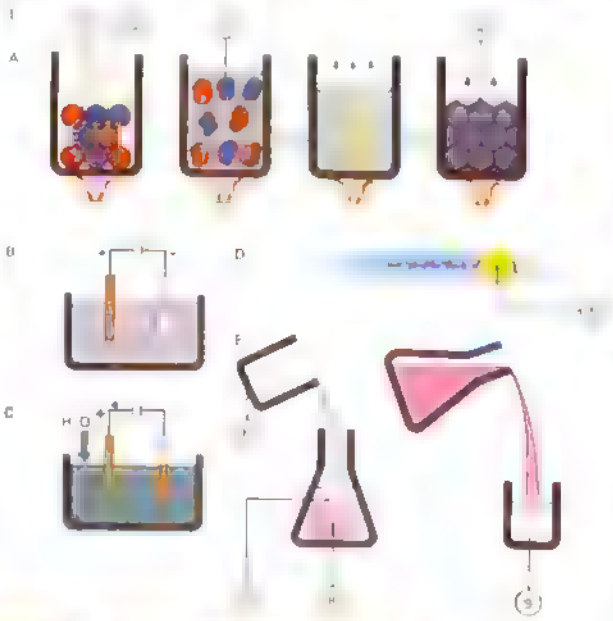
Kinds of reactions

Whereas certain reactions may occur spontaneously, they do not necessarily do so. A petrol/air mixture needs the spark to set it off because, in between the reactants and the products, there is a transition state which is of higher energy. It is necessary to give many reactants an energy "lift" to help them over the activation barrier. The larger the amount of energy arising from a reaction, the more molecules can be lifted over the barrier. And if the reaction happens fast enough, the result is an explosion.

The energy to initiate a reaction can come from a variety of sources. Heat is commonly

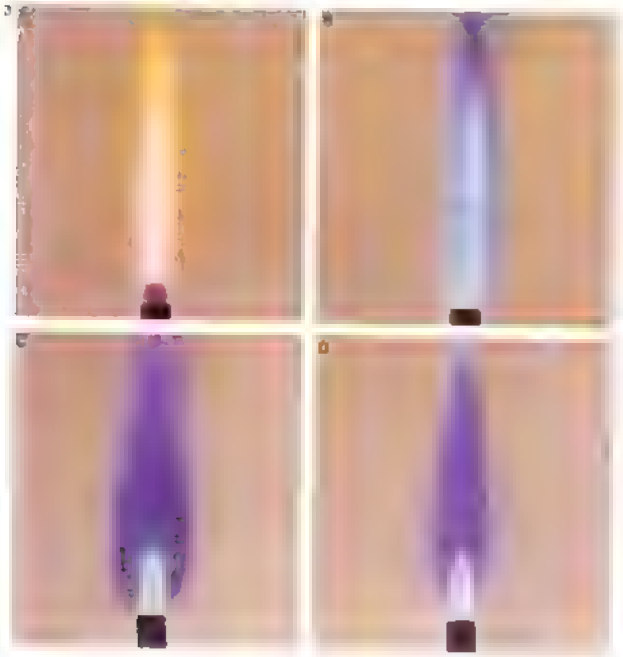
CONNECTIONS

See also

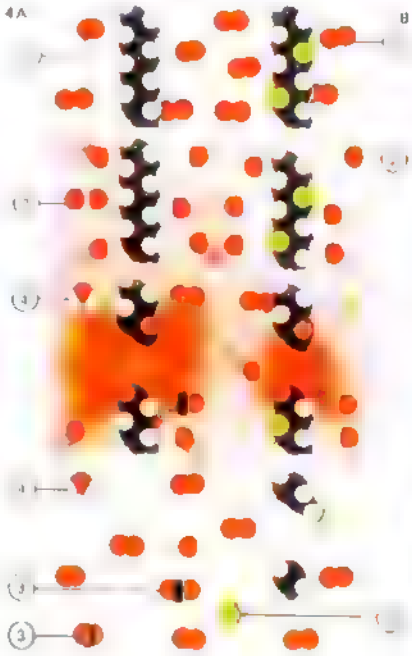


1 The application of heat [A] alone to aluminium sulphate [1] and potassium sulphate [2] produces no reaction, but they "combine" when dissolved in water [3]. If heating is continued to evaporation, alum [4] forms. Dry copper sulphate crystals [B] do not conduct current, but dissolve them in water [C] and electrolysis can proceed. Metals may react with a liquid [D]: a grain of sodium [5] dropped into a water bath melts, generating hydrogen. Solutions and other liquids react readily [E]. Phenolphthalein [6] added to a solution of alkali [7] produces a pink solution [8]. When this is added to an acid solution [9] the pink disappears.

2 In a chemical reaction, matter is neither created nor destroyed. This conservation of mass can be shown by a classic experiment in which a candle is burned inside a weighed bell jar [A]. At the end of the experiment, the weight of the jar and its contents [B] are the same as at the beginning, although a part of the candle – made up largely of carbon and hydrogen – has "disappeared" as volatile reaction products (water and carbon dioxide). It was only after scientists accepted the principle of conservation of mass in the late 1700s that a quantitative approach to chemistry became possible.



3 Unmixed ethylene burns with a luminous diffusion flame [A], reacting with oxygen drawn in from round the flame. If mixed with a little air, the ethylene gives a flame with three distinct layers: an inner cone of unburnt gas, a blue-green layer of reacting premixed gas and an outer cone where partially oxidized products of the premixed layer are burned by a diffusion flame [B]. Addition of nitric oxide to the mixture cuts down the amount of oxygen available for immediate combustion and the resultant flame [C] shows a complex series of reactions, as in B. But if more air is added to the gas mixture, the diffusion layer disappears [D].



4 A hydrocarbon fuel [A] will burn more readily than similar molecules in which some of the hydrogen atoms have been replaced by chlorine [B]. When mixed with oxygen [1] and subjected to a spark, both types of molecules burn, but the chlorinated one burns more slowly. The spark breaks oxygen molecules into reactive oxygen atoms [2]. These combine with carbon and hydrogen in the hydrocarbon to give carbon dioxide [3] and water [4]. Sufficient heat is produced to keep the reaction going rapidly. With the chlorinated material, more complex reactions take place more slowly. These produce, in addition, hydrogen chloride [5] but generate less heat.

used by chemists to help reactions along. But other forms of energy, such as light, are sometimes enough to initiate reactions.

There are a great many types of reactions but they can all be broken down into simple categories. There are reactions in which a single substance rearranges its chemical bonds to produce a different single substance (rearrangement), alternatively, it may break into two or more different parts (decomposition, fragmentation). Conversely two, or occasionally more, compounds can combine to form a single compound (addition). More often, there are a number of products.

Reaction requirements

When there is only one starting material, the activated state is achieved by molecules absorbing sufficient energy to initiate a change. For example, when the visual pigment rhodopsin in the retina of the eye absorbs light, one of its electrons is energized and the molecule is catapulted into a more energetic state. As a result, the molecule changes shape to form a different geometric isomer (substance whose molecules are made of the same numbers of the same atoms, but whose atoms are differently arranged).

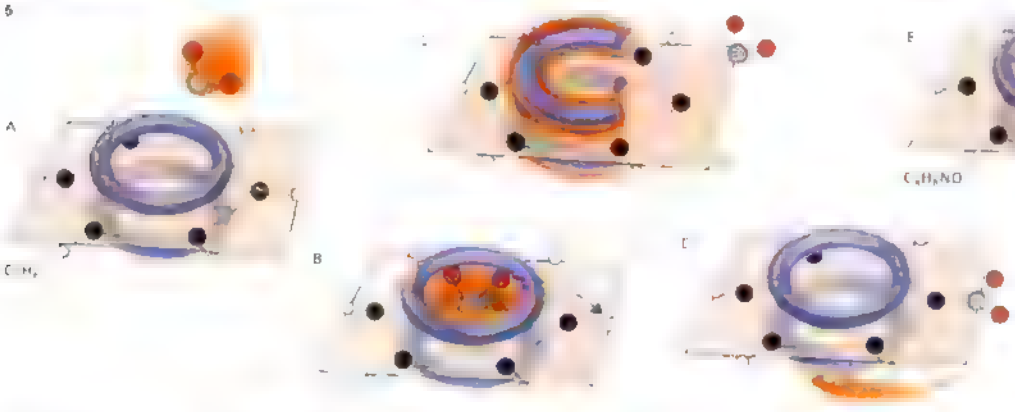
When different starting materials are involved, not only must they have enough energy to form the activated intermediate state, but they must also collide physically before the rearrangement of bonds can be completed. This is why some reactions take place under high pressure, as in a car engine because packing molecules more tightly together makes collisions more likely.

Another way to improve the chances of a successful collision is to provide a surface that has an appropriate geometry for bringing molecules together. Substances that provide such surfaces are among various types of catalysts. They speed up reactions by lowering the activation barrier. Other, negative catalysts slow or even prevent reactions.

To follow reactions as they occur and identify the various products, modern chemists use a large range of different techniques [6, 7]. Some of them, such as nuclear spin resonance and Mossbauer spectroscopy, are based on physical phenomena discovered during the past 30 or 40 years.

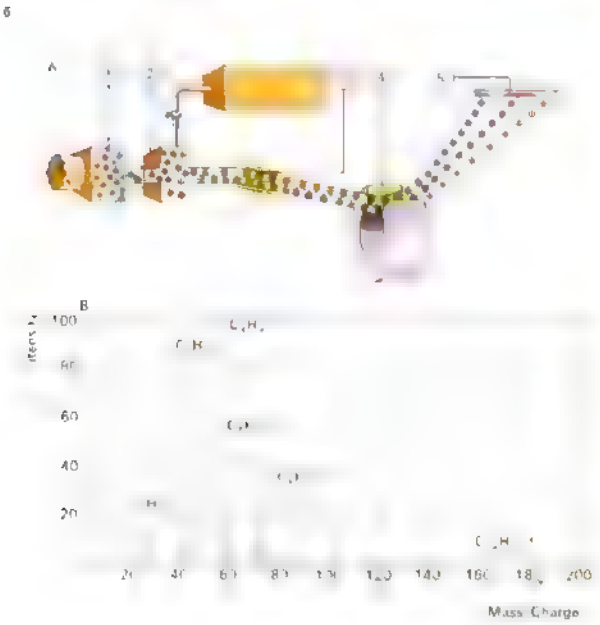
KEY

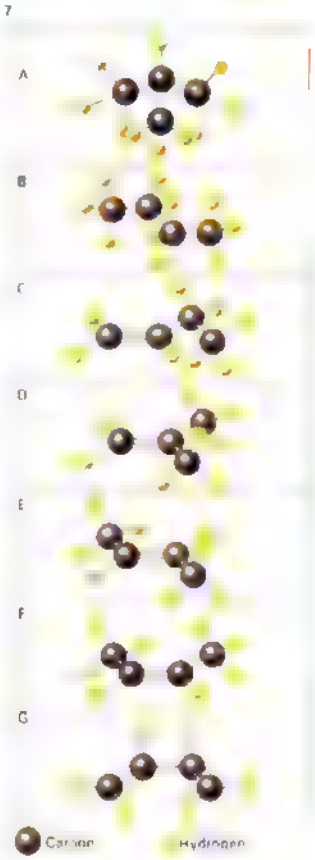
A reaction takes place as drops of alkali fall into a dilute solution of a copper salt. This is a precipitation reaction—one of many different kinds that can take place in chemistry. One of the chemist's major tasks lies in discovering how various materials react with each other.



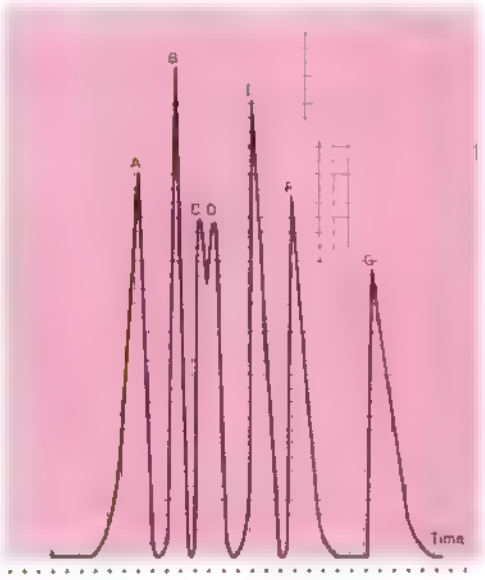
5 In a mass spectrometer (A), outer electrons of a compound are removed in an ionization chamber [1]. Positively charged ions pass into an adjacent chamber under vacuum [2] and are focused by electric [3] and magnetic [4] fields. The way they are deflected by these fields is characteristic for each ion, which can be identified by its position on a photographic plate [5]. A molecule such as n-dodecane breaks down into a number of fragments that produce various "peaks" on a graph (B). From the position of these, the parent molecules can be precisely identified.

6 The nitration of benzene produces a reaction that goes through several stages. Initially (A), the entering group approaches and associates weakly with the benzene ring (B). Then a rearrangement produces an unstable high energy intermediate (C) which breaks down to a complex (D) in which the leaving group is weakly associated with the ring. It ends with the departure of the leaving group (E).





7 Mixtures of compounds can be separated by gas chromatography. Gaseous molecules travel down columns of liquid-impregnated solids at different rates. Special detectors reproduce on a graph the peaks produced by the molecules as they leave the column. Each can frequently be identified by the time it takes to go through the column—the "retention" time. The chromatogram shows the separation of hydrocarbons: isobutane (A), n-butane (B), n-butene (C), isobutene (D), trans-but-2-ene (E), cis-but-2-ene (F), cis-1,3-butadiene (G).



Electrochemistry

Electrons are negatively charged particles that form a part of every atom and it is with the interactions of electrons from different atoms that chemistry is mainly concerned. An electric current is no more than a flow of electrons. Consequently, it is not surprising that electricity and chemistry are connected.

Early research

Studies of electricity and chemistry went hand in hand long before anyone knew of the existence of electrons. During the eighteenth century, there was much interest in static electricity, leading to the development of the Leyden jar (for storing "electric fluid" generated by friction) and the lightning conductor. However, it was not until the Italian physiologist Luigi Galvani (1737–98) found, towards the end of that century, that frogs' legs would contract if different metals were applied to nerve and muscle that current electricity was discovered. In 1795, another Italian, the physicist Alessandro Volta (1745–1827) [1], showed that this "animal electricity" could be produced without living tissue. He separated two pieces of metal by a cloth moistened with salt solution and thus made the first electrical battery. Within five years, it was discovered in England that current from such a battery could decompose water into hydrogen and oxygen gases. Thus, the foundations of electrochemistry were laid. By a chemical reaction involving two metals, a flow of electrons can be produced, such flow can bring about other reactions.

Batteries soon became important equipment in every laboratory and led to many new discoveries, such as the isolation of the elements sodium and potassium in the first decade of the nineteenth century by Humphry Davy (1778–1829).

Chemical reactions

When a metal such as zinc forms compounds, it does so in many instances by losing two electrons to form a doubly positive zinc ion (Zn^{2+}). Metals differ in the ease with which they lose electrons, so that if a piece of zinc metal is placed in a solution of copper sulphate (which contains Cu^{2+} ions), the zinc gives up electrons to the copper. The net result is that zinc is converted to zinc sulphate, which contains Zn^{2+} ions, and copper ions become metallic copper.

When an element gains electrons to form a negatively charged ion, it is said to be reduced; if it loses electrons to form a positively charged ion, it is said to be oxidized. A reaction where reduction and oxidation cancel each other out, as in the zinc/copper sulphate case, is called a redox reaction. Redox reactions can be tapped to supply electric currents by preventing the reduction and oxidation from occurring at the same place. A battery can be made by suspending zinc in zinc sulphate and copper in copper sulphate and linking the two solutions by a porous partition and the metals with wire.

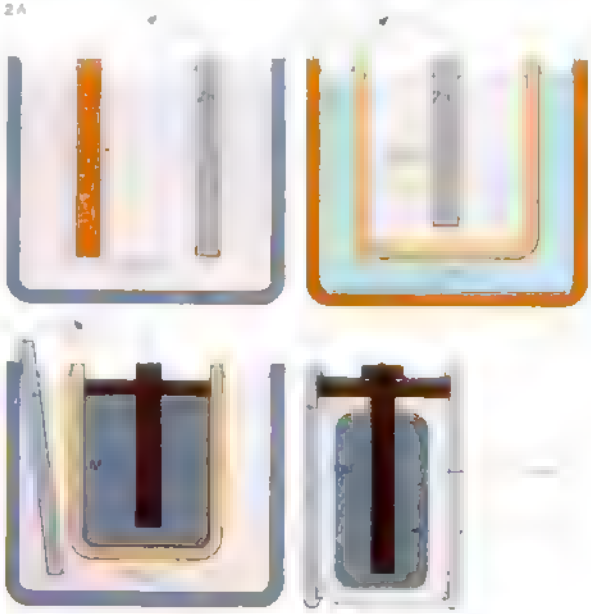
Each of the reactions in such an arrangement is called a "half cell". When two "half cells" are added together, a cell is completed and the voltage it produces depends on the particular half cells that make it up.

That different batteries produce different, but specific, voltages depending on their chemical composition is not surprising, in view of the differences in reactivity between different elements. The reverse also seems

CONNECTIONS

See also

1 Alessandro Volta, professor of natural philosophy at the University of Pavia, Italy, constructed in 1800 an "artificial electrical organ", an apparatus he described as like the electrical organ of the electric eel. Made by piling alternate discs of copper and zinc, each pair separated by a piece of brine soaked cloth, his electrical organ was one of the first scientific batteries ever made.



2 The voltaic cell (A) consists of a jar containing sulphuric acid (H_2SO_4) in which are suspended a copper (Cu) anode and a zinc (Zn) cathode. When the circuit is closed a current flows, zinc ions pass into the acid from the cathode and hydrogen (H_2) is deposited on the anode. Hydrogen eventually obstructs the reaction, but not so in the Daniell cell (B), in which the zinc cathode and sulphuric acid are encased in a porous pot surrounded by a solution of copper sulphate ($CuSO_4$). The copper case acts as the anode. The Leclanche (C) and dry cells (D) have carbon (C) anodes enclosed in manganese dioxide (MnO_2) in ammonium chloride (NH_4Cl), both types have zinc cathodes.

3 Heart pacemakers like the one shown here, and miniature hearing aids, can be powered by batteries. These are examples of primary cells. Secondary cells or storage batteries can be recharged. Early batteries all had metallic plates separated by solutions of salt-like chemicals. The dry, or Leclanché, cell replaced the liquid with a paste. Such batteries, which use zinc and carbon (with a manganese dioxide depolarizer) as electrodes, are the type used in transistor radios, torches, and many other everyday appliances. In recent years, battery technology has led to the production of very small but highly reliable batteries which also use zinc and a metal oxide, in this case mercury.

4 Elements which gain or lose electrons easily are often prepared from ionic compounds by electrolysis. Chlorine [1], for example, can be produced by electrolysis of a sodium chloride solution [2] at graphite anodes [3]. Hydrogen [4] from the electrolysis of water is released at the cathodes [5]. Left behind in solution are sodium and hydroxyl ions, giving a solution of sodium hydroxide [6]. In the practical cell a partition keeps the products (hydrogen and chlorine) apart to prevent them from reacting together to produce hydrogen chloride. Similarly a diaphragm keeps the hydroxyl ions away from the chlorine to prevent them from reacting to form sodium hypochlorite.

logical: that a particular quantity of electrons should produce a particular amount of change in a substance. The quantitative relationships between electricity and chemical reactions were stated during the nineteenth century by Michael Faraday [Key]. The extraction and electroplating of metals, and the production of reactive electronegative elements such as chlorine [4] and fluorine, are often done electrolytically

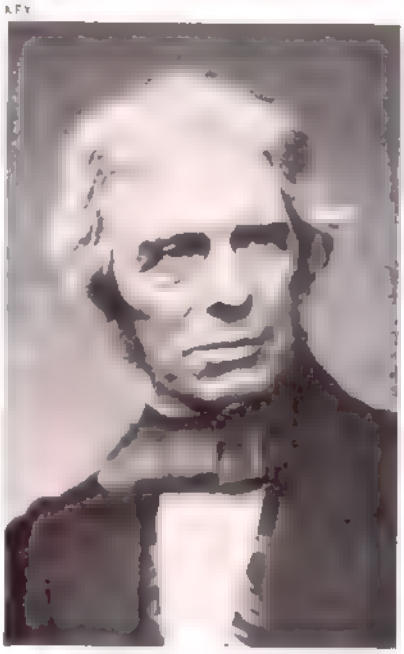
Electrolysis

The products from electrolysis reactions [2] sometimes depend on the conditions used, as well as the amount of electricity. If fused (molten) sodium chloride is electrolyzed, sodium metal forms at one electrode and chlorine gas at the other However, when a solution of sodium chloride is electrolyzed, using a graphite anode (positive electrode) and an iron cathode (negative electrode), chlorine and hydrogen gases are produced, leaving behind sodium hydroxide solution

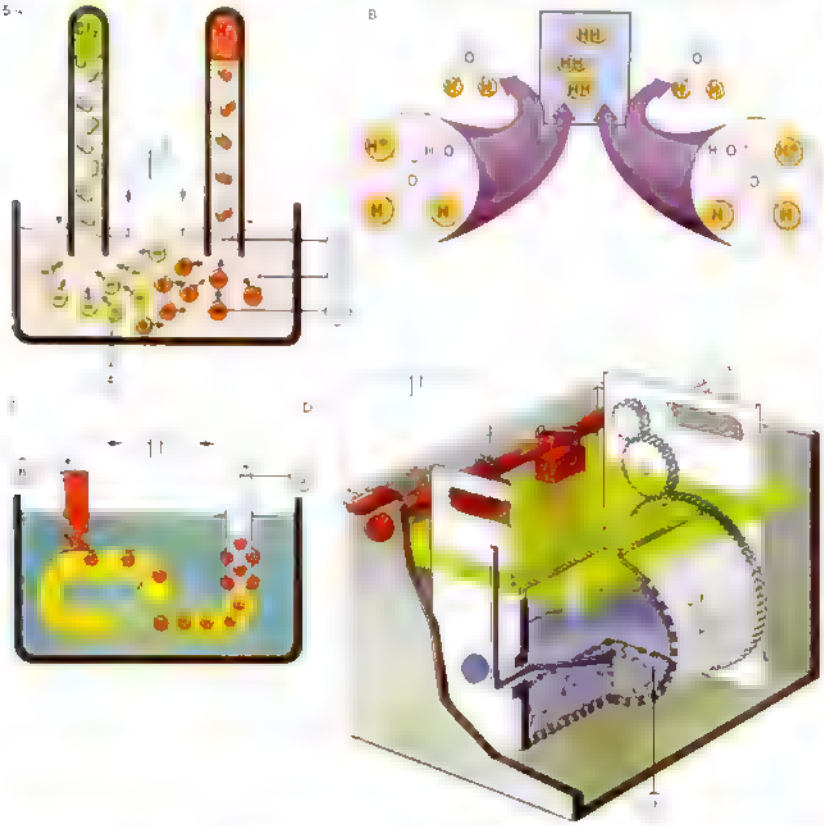
The ions of different elements may be positive (cations) or negative (anions). In a solution that is being electrolyzed, cations are attracted to the cathode and anions to the anode If aluminium is made into an anode in an acid solution, a very thin layer of aluminium oxide forms on it. This anodization protects the aluminium from corrosion and is used on a wide range of articles.

Many oxidation reactions are used in everyday life for example, the burning of petrol in a car is such a reaction. Instead of releasing the energy from such a reaction as heat, it can be converted into a flow of electrons in a "fuel cell" These cells are theoretically much more efficient energy converters than heat engines. However, difficulties in designing suitable fuel cells for everyday purposes have meant that their use has been limited largely to applications where cost is not an important factor

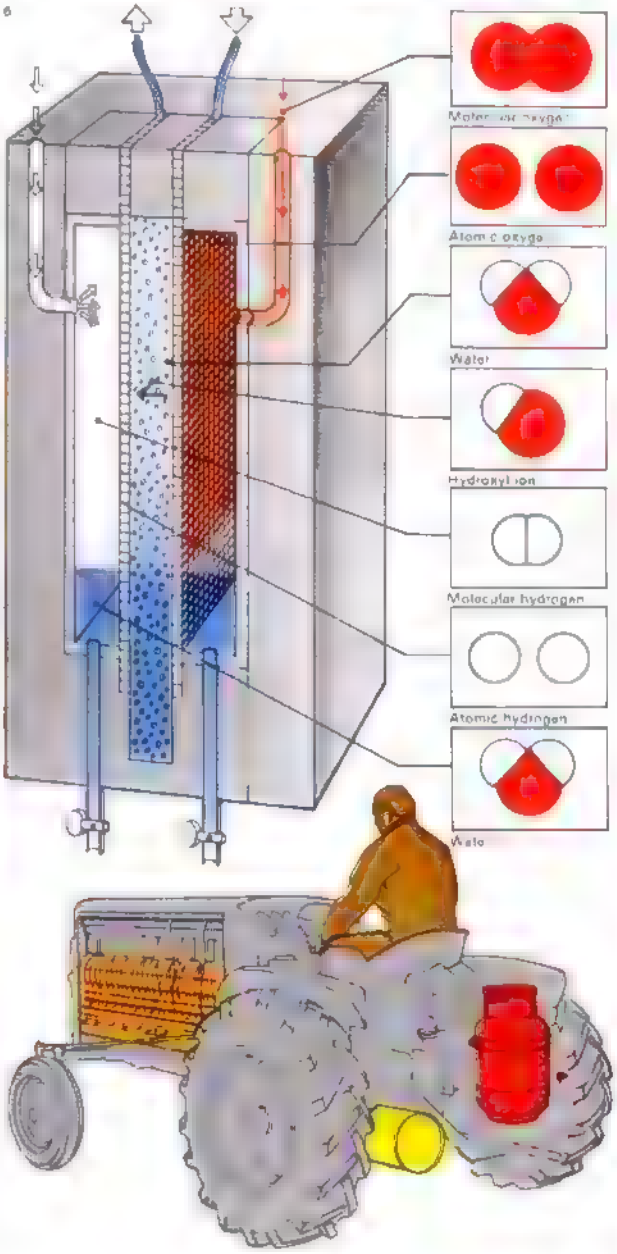
The chief commercial application of electrochemistry is electroplating by means of electrolysis. For example, decorative metals such as gold and silver are electroplated onto articles of jewellery, whereas chromium is electroplated onto steel (preferably over base layers of copper and nickel) to provide resistance to corrosion



As a youth, Michael Faraday (1791–1867) attended Humphry Davy's lectures at the Royal Institution in London He copied these out and sent them to Davy with a request for employment. From Davy's assistant, he rose to become professor of chemistry at the Royal Institution, a post he held for more than 30 years Most of Faraday's work was in physics – particularly in the field of electromagnetic induction He also founded the science of electrochemistry through his discovery of the quantitative relationships between the amount of electricity passed through a solution and the amount of substances deposited as a result



5 Electrolysis can easily be shown A current is passed [A, between platinum electrodes [1] through a dilute hydrochloric acid electrolyte [2] Positively charged cations [3] move towards the negative electrode (cathode) and anions [4] move to the positive electrode (anode) The hydrogen ions combine with water to form hydronium ions, H_3O^+; when two hydronium ions [B] reach the cathode, they each receive an electron, thus forming atoms of hydrogen which combine to form a molecule of gas. In copper plating [C], copper from the copper sulphate electrolyte is deposited onto the object to be plated [5], while copper from the anode [6] is drawn into solution. A modern barrel electroplating machine [D] can plate many small objects [7] simultaneously



6 The fuel cell, like a battery, uses the energy generated during a chemical reaction to produce electrical power. One simple cell uses oxygen and hydrogen as "fuels" to produce electricity, the water formed being a by product The electrolyte in this case is located in a very thin, water-saturated membrane This allows ions to pass through it, but does not allow the passage of atoms or molecules. The electrodes are of wire mesh, coated with platinum Molecular hydrogen and oxygen are fed to them from gas chambers The cathodic platinum converts oxygen to hydroxyl ions, which move across the membrane and react with hydrogen at the anode to form water Electrons released by the hydrogen traverse the external circuit as an electric current to help form hydroxyl at the anode. A battery of such cells can power a tractor, with fuel tanks holding liquid oxygen and hydrogen either as a liquid or as a solid hydride

Chemical analysis

One of the main branches of chemistry – chemical analysis – is concerned with determining the composition of a substance or a mixture of substances. Identifying the ingredients is termed qualitative analysis, whereas determining their precise proportions is called quantitative analysis. Organic chemicals (the large class of compounds containing the element carbon) and inorganic chemicals (all other compounds) require different analytical techniques in the laboratory

Methods of inorganic analysis

Qualitative inorganic analysis is generally carried out on a semi-micro scale [1] using small quantities – much less than a gramme Chemists make preliminary tests on a dry sample of a substance and these give general information about its composition The effect of heat may cause a colour change, sublimation or the evolution of a gas. A reagent is added to a solution of the sample and the resulting mixture examined for the evolution of a gas, a precipitate or a coloration

Metal ions are identified by a systematic separation into "groups" A variety of techniques exists, but metals are usually split into groups by adding a series of reagents and collecting any precipitate produced Group precipitates are, in turn, separated and identified by characteristic reactions

Quantitative inorganic analysis may be carried out using either volumetric or gravimetric methods. Volumetric analysis involves reacting a solution of known concentration, referred to as a "standard" solution, with a solution of the substance to be determined. After preparing the standard solution the chemist carries out a titration [Key] in which one solution is slowly added to the other. From the concentration and volume of the standard solution, the concentration of the "unknown" can be calculated

Gravimetric analysis involves preparing a solution with a known mass of the sample This solution is then reacted with a chosen reagent so that the desired component is completely separated, generally as a precipitate. The product, which must be pure, is isolated and weighed and the amount of the component calculated

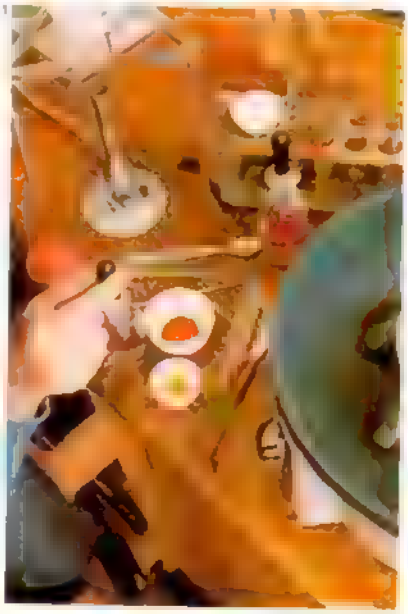
Merely identifying the elements present is insufficient to make a definite description of an organic compound. The ability of carbon compounds to exhibit isomerism (in which two different substances can contain the same chemical elements, in the same proportions, but combined in a different way) means that the arrangements of the elements present must also be analytically determined

Methods of organic analysis

Identification of the elements in an organic component involves systematic elimination of all possible elements one by one. Carbon and hydrogen are always present and tests for them are rarely carried out But tests are made to identify other elements. An example is the Lassaigne sodium fusion, which reveals the presence of nitrogen, the halogens (chlorine, bromine and iodine) and sulphur A knowledge of the elements allows a chemist to allocate the compound to a main group. This is followed by the application within the group, of classification tests for functional groups, which determines the types of organic compound present

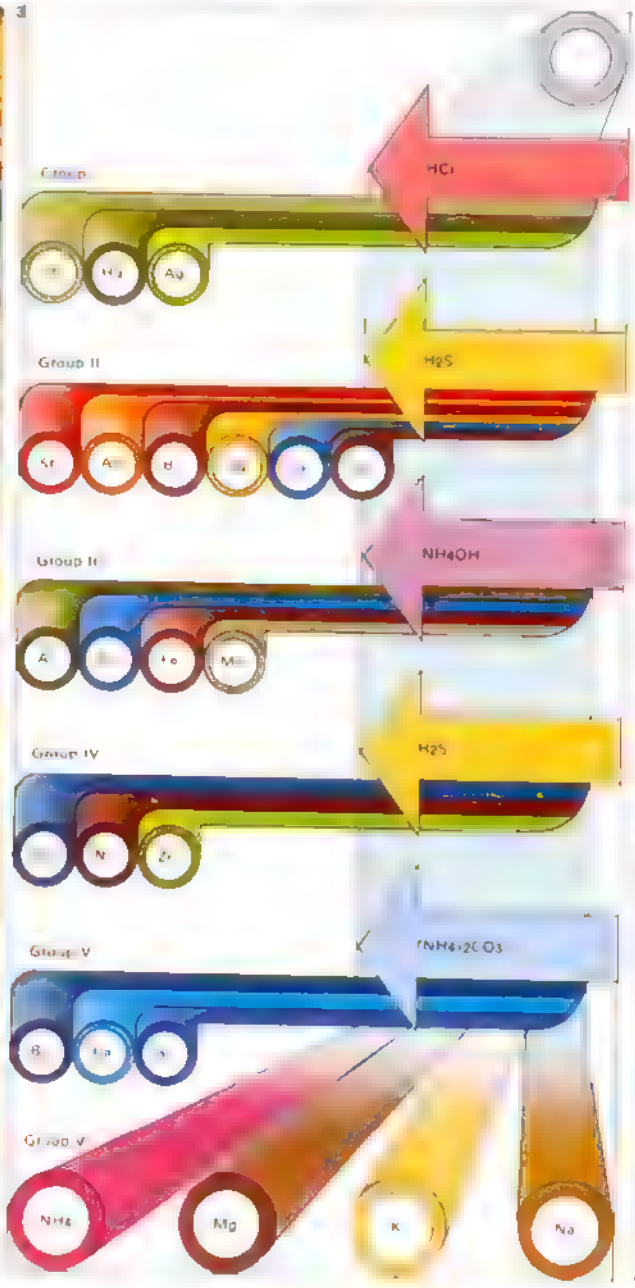
CONNECTIONS

See also
Classification of



1 Analysis on a semi-micro scale saves time and money Small test tubes are used for reactions and tapered tubes are used for the centrifuge Solutions do not mix well in these small tubes so a stirring rod is necessary Solutions are handled in a teat pipette and each reagent bottle should be fitted with its own "dropper" Solids require the use of a semi micro spatula To avoid the dangers of "bumping" during heating, solutions are heated indirectly in a metal block, a though evaporation to dryness requires the use of a small crucible Identification of gases is usually carried out with a bubble-cap fitted to a test tube

2 The melting-point is a characteristic of an organic compound and can be measured using this automatic apparatus, to help to identify it. A sample is first heated quickly to get an approximate value and then slowly melted to obtain a more accurate reading



3 Metal ions can be detected in qualitative chemical analysis by a systematic separation into groups by means of a series of characteristic precipitation reactions The chemist tries to dissolve a sample of the substance to be analysed in dilute hydrochloric acid Metals with insoluble chlorides lead, mercury and silver – constitute Group I. Next hydrogen sulphide is bubbled through the acid solution, a sulphide precipitate indicates one of Group II metals – antimony, arsenic, bismuth, cadmium, copper or tin The addition of ammonia solution precipitates the hydroxides of the Group III metals aluminium, chromium, iron or manganese The Group IV metals have sulphides that are precipitated from alkaline solution by bubbling in hydrogen sulphide gas they are cobalt, nickel and zinc The addition of ammonium carbonate to the remaining solution at this stage of the analysis precipitates the carbonates of barium, calcium or strontium, the Group V metals Group VI of the analysis table contains the metals magnesium, sodium and potassium as well as the "metallic" ion ammonium, left after eliminating all other possible metal ions This analysis scheme can be carried out on the semi-micro scale and it can be enlarged to include some of the less common metals It reveals only the presence of a metal and not its quantity

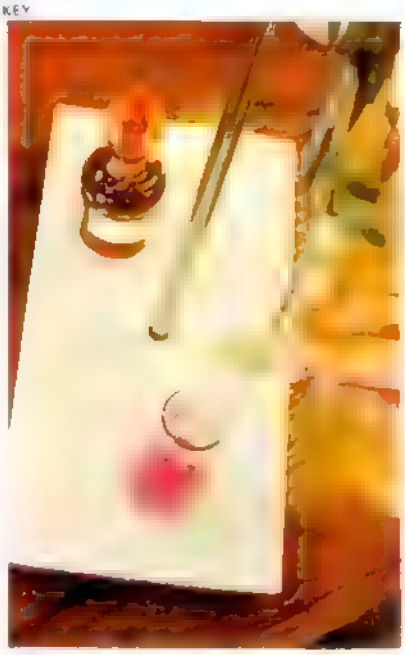
Quantitative organic analysis also involves estimation of the elements present, followed by purification and a determination of molecular weight to give the empirical and molecular formulae. The amounts of carbon and hydrogen are found by completely oxidizing a known mass of the organic compound and weighing the carbon dioxide and water formed. Then any other elements are estimated by a variety of methods.

These results allow the chemist to calculate the percentage composition of the substance (the proportions of each element present) and to determine its empirical formula. The molecular formula is found by comparing the empirical formula with the molecular weight. Dissolving a sample of the substance in a solvent affects the physical properties of the solvent. The lowering of vapour pressure, the elevation of boiling point and the depression of freezing point are all proportional to the mole fraction (concentration in terms of molecular weight) of the dissolved substance present in the solution. The concentration also affects the osmotic pressure, viscosity and light scattering properties of solutions. Careful measurement of one of these effects is then followed by a calculation of the molecular weight before the chemist has enough information to complete the analysis.

Modern instrumental analysis

Various forms of chromatography are based on the fact that different substances diffuse or are absorbed at different rates. Spectroscopy [7] makes use of the fact that each species of atom has a unique characteristic spectrum. A spectrum is produced when atoms, ions or molecules are excited by absorbing energy and may be observed by using a prism or, preferably, a diffraction grating.

In mass spectroscopy [5] a substance is bombarded with low-energy electrons and fragmentation produces a number of positive ions. Ions of the same mass/charge ratio are focused by magnetic or electrostatic fields and detected photographically or electronically. The highest mass/charge ratio can give the molecular weight. Investigation of the fragmentation pattern determines the molecular structure.

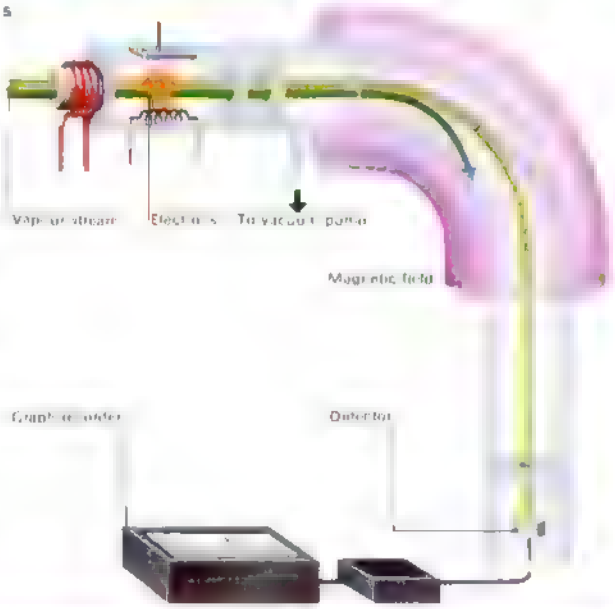


A titration is performed to estimate the unknown concentration of a solution by reacting it with a "standard" solution of known concentration. This is usually added from a burette to a fixed volume of the "unknown". The end point of the reaction is shown either by a visible change in the reactants or by the addition of a chemical indicator. Acid/base indicators have different colours according to the hydrogen-ion concentration (pH) of the solution and change as the pH of the solution changes. The pH at which colour changes occur varies so an indicator can be selected that shows a colour change at a pH close to the end point.

4

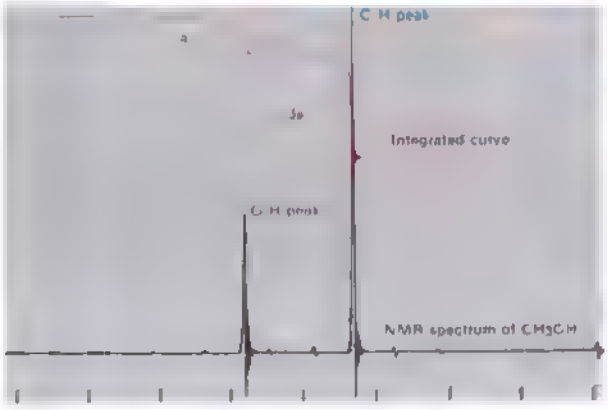
4 Molecular weight may be determined by the depression of freezing-point. When a substance is dissolved in a solvent the freezing point is depressed. If dilute solutions are used then the depression is directly proportional to the number of molecules of the solute in unit mass of the solvent. The molecular depression constant of a solvent is the depression of freezing point produced for one mole of solute in 100g (3.5oz) of solvent. Experimentally this quantity would need too high a concentration so dilute solutions are used and the constant calculated by proportion. The most convenient freezing agent is ice and water.

5



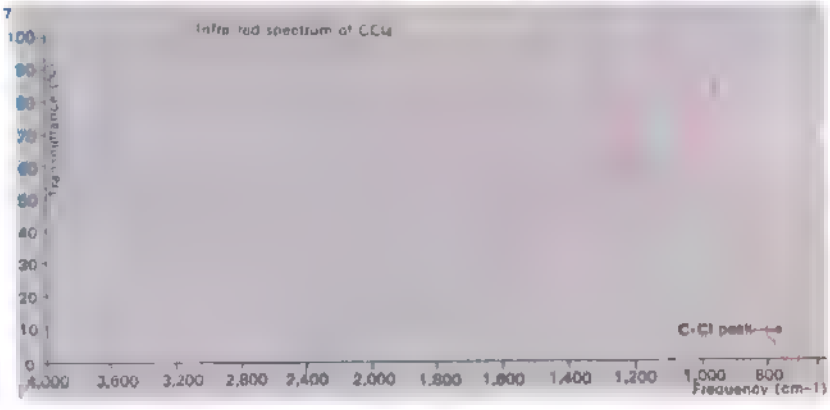
5 The masses of atoms can be compared with great accuracy using a mass spectrometer. A vaporized sample is ionized by electron bombardment and the beam of positive ions produced is accelerated to a constant speed. Application of a strong magnetic field deflects the beam – the lightest ions being deflected most and the heaviest least. The field is adjusted so that ions of a particular mass fall on to a detector, either a moving photographic plate or an electrometer. From an electrometer the signal is amplified and recorded on a graph. Further adjustment of the magnetic field allows ions of different mass to be recorded.

6



6 Nuclear magnetic resonance (NMR) spectroscopy of a substance with a molecular formula of CH_3O shows two peaks. The areas under the peaks are integrated automatically and indicated by the upper curve. This shows a ratio of one to three. The inference is that the four hydrogen atoms in the molecule are arranged so that three are in the same environment, and the other is different. The larger peak is produced by those in the $O\text{-}CH_3$ group and the smaller by that in the O-H group – indicating the structure CH_3OH methanol. Normally a "standard" is added and all lines are measured relative to it. Tetramethylsilane or a sodium salt of 4,4 dimethyl,4,silapentanesulphonic acid is a suitable standard.

7



7 Infra-red radiation is absorbed by the chemical bonds in an organic compound and can cause vibrational effects in them. The frequency of the absorbed radiation is characteristic of the bond concerned so that measuring these frequencies provides a means of determining the bonds present and analysing the compound. This is the basis of infra-red spectroscopy. The absorption frequencies are measured electronically and plotted as a series of peaks on a graph. In this example the main absorption peak at a frequency of 750cm^{-1} is due to stretching of a carbon chlorine bond, indicating that the substance producing it is probably tetrachloromethane (carbon tetrachloride, CCl_4). The minor peak at 1,550cm^{-1} is probably a harmonic of the main one. The bending and rocking of chemical bonds after infra red absorption also produce characteristic peaks on the spectrograph and aid the analysis.

Towards the chemistry of life

Only half a dozen of the 93 or more chemical elements that occur naturally on earth make up the bulk of living matter, and life's diversity is due largely to the combining properties of just one element: carbon. Carbon atoms can form chemical bonds with each other to produce an extensive range of basic structures. These can be modified by the addition of the atoms of other common elements of life – hydrogen, oxygen, nitrogen, phosphorus and sulphur – to produce the enormous diversity of chemical substances found throughout the living world.

Isomers and polymers

Many naturally occurring carbon compounds have another distinctive property. A single atom of carbon can form chemical bonds to four different atomic groupings. These groupings can be arranged to produce two different molecules that are as similar to and as different from each other as a pair of gloves [1]. They are called optical isomers and where two of these are possible, usually only one form occurs naturally

Many key substances in living organisms are polymers – giant molecules containing thousands or even millions of individual atoms linked together. Carbohydrates, proteins and nucleic acids are all polymers. But they are all made by joining together small molecular building blocks rather than individual atoms.

Carbohydrates, for example, are all made from small molecules called sugars, or saccharides. Common table sugar is not one of the simplest: it is made by linking two smaller saccharides. Chemists call it a disaccharide It is an ingredient of many proprietary foods, such as pickles and ketchup [3]

Like all carbohydrates, saccharides are composed of carbon, hydrogen and oxygen. These elements are generally linked together in such a way that a loop is formed between the ends of the molecule by an oxygen atom bridging two carbon atoms

Sugars and fats

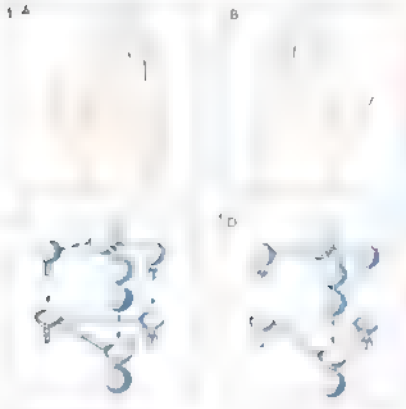
Sugars [2] not only supply living organisms with energy, but they also make up a broad range of polymeric substances, such as starch and cellulose Starch, for example, is the chief carbohydrate in potatoes, rice and bread [3] Human beings cannot digest cellulose, but the polymer can be broken down chemically to form molecules of glucose, which is an example of a simple sugar (monosaccharide) Sugar from beet or cane is sucrose

Virtually all food has sugars or polysaccharides in it Almost as common in food are the lipids, composed solely of carbon, hydrogen and oxygen. Lipid is a general term that includes oils, fats and waxes, all of which have similar chemical structures. In a simple lipid, the same sort of carbon-oxygen-carbon bond that holds a disaccharide together joins three fatty acids to a small molecule called glycerol (also known as glycerine) [4B] The resultant triglyceride may be a liquid (oil) or solid (fat) at room temperature, depending on the structures of the fatty acids, which can all be the same (as in olive oil) or all different. All meats contain some fat [3]

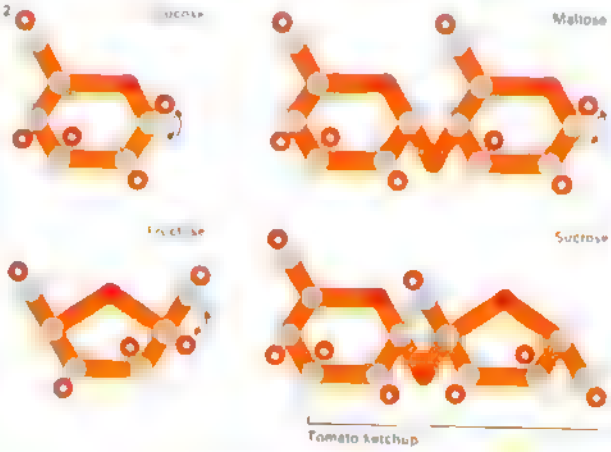
The so-called unsaturated fatty acids contain carbon-carbon double bonds. In recent years, margarines containing them have been available in many countries because a link has been suggested between high consump-

CONNECTIONS

See also

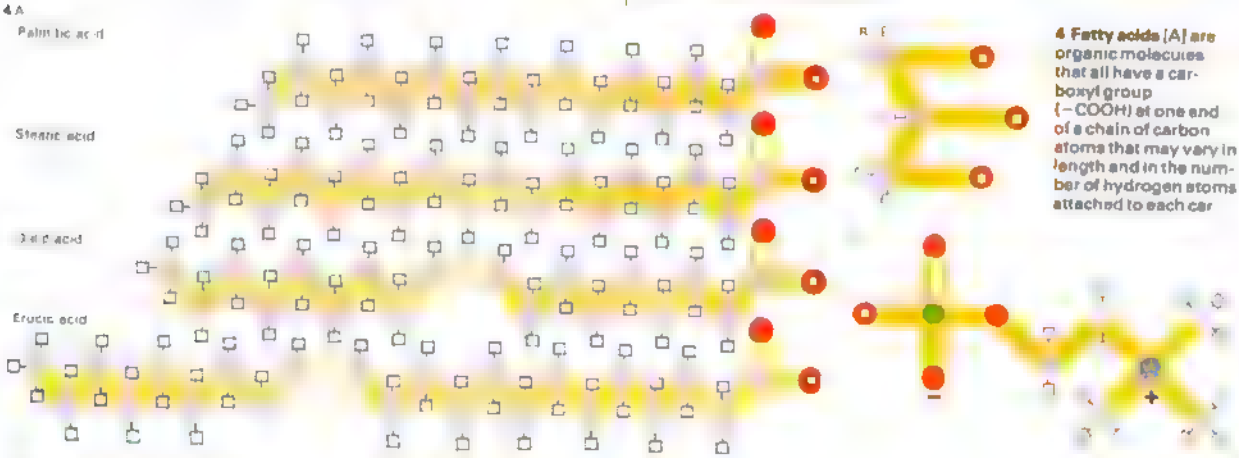


1 Most biologically active molecules depend on a particular shape for their function because they interact with other molecules and must fit them like a key in a lock [C] Because of the ways in which molecules are constructed, it is possible for many of them to exist in more than one form This variation in form is called isomerism. The most important type of isomerism in biological chemistry is called optical isomerism, in which molecules differ only in so far as they are mirror images of one another (just as a left and right hand [A, B] are mirror images). The upper molecule in C fits the lower one But the upper molecule in D the mirror image of that in C does not fit When a chemical is made during biological processes, the product is nearly always 100 per cent of one particular isomer



2 Among the simplest of life's molecules are the saccharides, or sugars They are made only of carbon, hydrogen and oxygen atoms, the hydrogens and oxygens almost always in a 2 1 ratio, as in water They therefore became known as carbohydrates (hydrates of carbon) The simplest sugars, or monosaccharides, are the pentoses and hexoses, with 5 or 6 carbon atoms. Often the carbon atoms join together to form a ring, as in glucose and fructose Glucose molecules can link together to form medium-sized molecules, such as maltose, an important sugar in bread and beer making, or very large molecules such as starch When monosaccharides link together, they need not all be of the same type Common table sugar, or sucrose is made by linking one molecule of glucose with one of fructose

3 A meat and salad roll contains protein, carbohydrate and fat



4 Fatty acids [A] are organic molecules that all have a carboxyl group (–COOH) at one end of a chain of carbon atoms that may vary in length and in the number of hydrogen atoms attached to each carbon. They are a fundamental component of lipids, compounds found in materials as diverse as bacon fat and olive oil In the most common lipids, three fatty acid molecules are linked chemically to one molecule of glycerol [B] In some important biological molecules, one of the fatty acids may be replaced by a phosphorus-containing molecule, such as choline phosphate [C], to form a type of compound known as a phospholipid

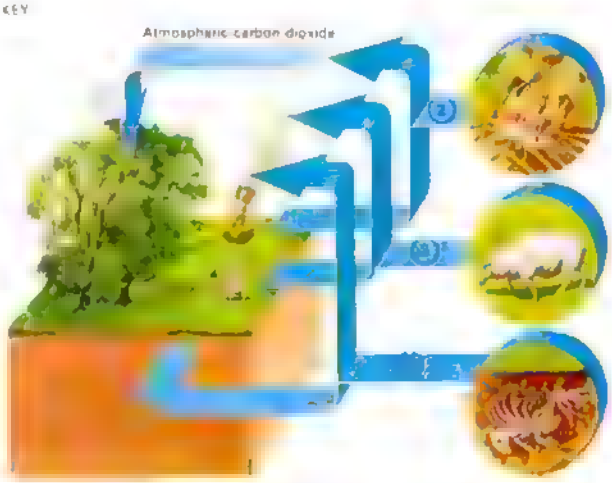
tion of saturated fats (such as butter) and the incidence of heart disease. Lipids also provide energy for animal cells but, in more complex forms, they play other roles, such as "insulators" for nerve fibres.

Triglycerides are the simplest lipids. An important and more complicated example is cholesterol, widely found in dairy foods such as cream and cheese [3]. It is a major constituent of gall-stones and also implicated in heart disease. Cholesterol, with a complex chemical structure related to that of human sex hormones, is a type of molecule called a steroid. Steroids may be synthesized in living systems from a molecule called squalene, a kind of terpene. Some terpenes are made from only carbon and hydrogen, but others also contain oxygen and nitrogen. They include not only substances such as turpentine, vitamin A and cholesterol, but many flavours and fragrances as well.

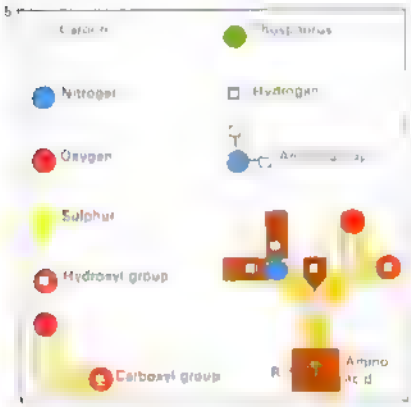
Amino acids

From small groupings of carbon, oxygen, nitrogen, hydrogen and occasionally sulphur, all joined in a particular pattern, come the amino acids [5], the building blocks for proteins. Proteins are major constituents of some of the most important foodstuffs, such as meat, fish and eggs. They are usually made from only 20 different amino acids and yet they have a wide range of valuable properties. In addition a number of other, less common amino acids are known. These can combine to form compounds such as the antibiotic valinomycin, which is made by linking together a small number of amino acids, and the extremely poisonous seven and eight amino-acid rings occurring naturally in toadstools of the genus *Amanita*, such as the death cap and fly agaric.

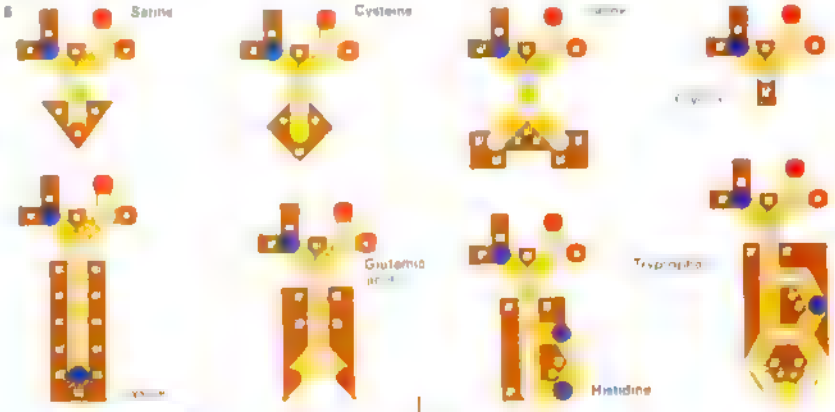
More complex than the amino acids are the nucleotides, the basic building blocks of the nucleic acids, which carry genetic information. Each nucleotide is made from one of the five types of base [7] which is joined to a sugar and this, in turn, is joined to a phosphate grouping. Some foods are very rich in nucleic acids and too much of them can cause an illness, such as gout, in which the mechanism that usually deals with the bases falters or fails completely from overwork.



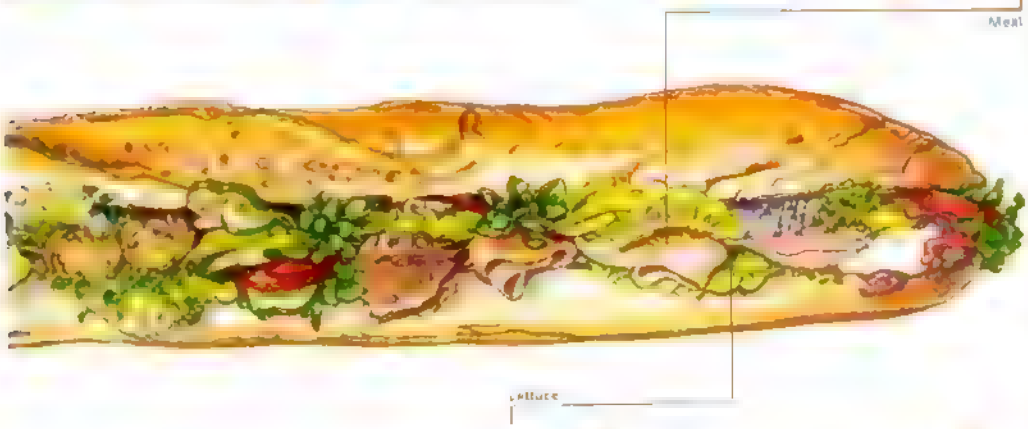
Carbon, the basic element of organic chemistry, undergoes a natural cycle in the environment. It exists in the form of carbon dioxide in the atmosphere. From there it is absorbed by plants [1] to build carbohydrates in green leaves. When plants burn [2], and animals breathe out [3], carbon dioxide passes back into the air. Also in decay ing plant and animal remains [4], carbohydrates are broken down to release carbon dioxide into the atmosphere.

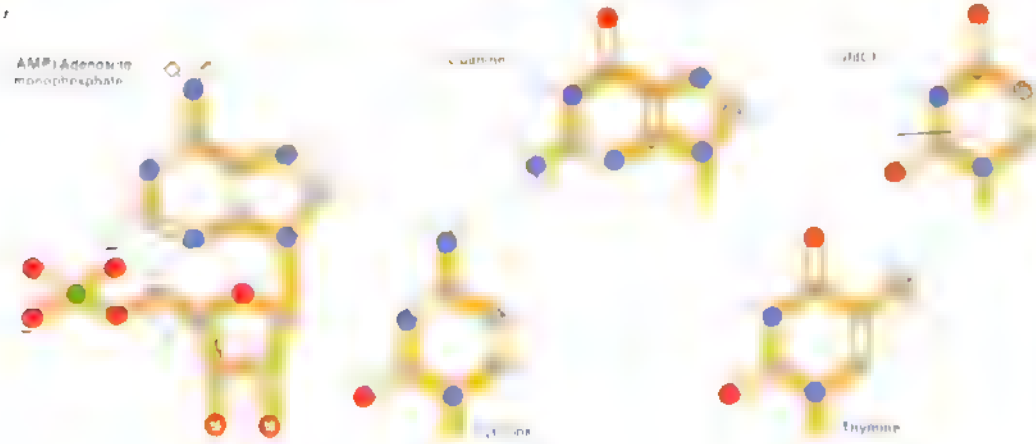


5 About 20 amino acids play essential roles in the structure and function of living organisms. All contain carbon, hydrogen, nitrogen and oxygen atoms, a few also contain sulphur. All except proline have a free unsubstituted amino group and a free carboxyl group. In addition, each amino acid has a characteristic "R" group attached to this carbon atom. This key applies to both of these pages.



6 Amino acids usually occur in living organisms linked together in hundreds to form complex molecules. But monosodium glutamate, a simple derivative of glutamic acid, has been used as a food additive to enhance the flavour of meat. The "R" group attached to the central carbon atom in an amino acid can vary widely – from the simple hydrogen atom in glycine, to the complex groups found in the other amino acids illustrated here. In proline, the end of the "R" group is linked also to the nitrogen atom of the amino group, so looping round on itself to form a ring structure.





7 All living matter, be it in the form of liver or lettuce, contains substances called nucleic acids. These are the giant molecules that ultimately are in control of all living processes. Protein made by bacteria, a possible source of synthetic food for man, is rich in nucleic acids. But an excess of these substances or foods containing them can produce unpleasant side-effects in human beings. They are formed by linking together large numbers of a few simpler molecules called nucleotides. A single nucleotide is, in turn, made by linking together three even simpler chemical groupings: a phosphate, which is joined to a 5-carbon monosaccharide (either ribose or deoxyribose), to which a nitrogen-containing base is attached. Only five of these bases are common: adenosine (shown here linked to phosphate), cytosine, guanine, thymine and uracil. The nucleic acids DNA and RNA contain chains of four such bases linked through a sugar and phosphate.

The chemistry of life: biochemistry

All living things – plants and animals – build up and break down different chemicals. These chemical processes ensure that an organism has an adequate supply of both the basic materials and the energy it needs for survival. A person eating a salad roll derives energy from it, which he may use up by running several kilometres, but there is no obvious, direct link between the salad roll and the exercise. Biochemistry, through interlocking reactions, provides that link.

Fundamentals of biochemistry

The complexities of biochemistry can be reduced to two fundamental processes. The first is the way in which living cells develop an energy currency. This, like ordinary money, can be used to exchange one vital commodity for another. The second is the use of substances called enzymes [2] as go-betweens to reduce the amount of energy needed to make many chemical reactions essential to life take place fast enough.

The "currency" used by living cells is a chemical called adenosine triphosphate (ATP) [1]. Closely related to one of the units from which nucleic acids are built, ATP can break down to form adenosine diphosphate (ADP) and phosphate. In doing so, it can supply energy for a biochemical reaction – either one in which simple molecules are built up into more complex ones, or one that controls an activity such as muscle contraction [Key]. On the other hand, where a biochemical reaction gives off energy, as in the breakdown of sugars, that energy can be used to reform ATP. Consequently, an organism can balance energy inputs against energy outputs by recycling ATP.

Although some biochemical reactions ultimately give off energy and others use it, all need an energy push to get them started. The strength of this push can be decreased if the reacting molecules are close together and are lined up in the correct relationship to each other. Substances called catalysts introduce reacting molecules to one another more efficiently and so reduce the amount of energy needed to get the reaction moving. The overall effect is to make chemical reactions proceed much more quickly.

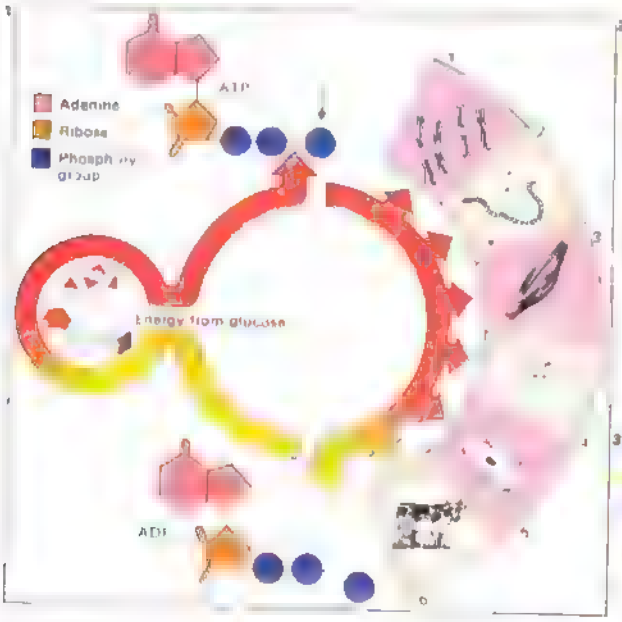
Life relies heavily on a special class of catalysts, the enzymes [4]. These are made mainly from protein, but they may also include metal atoms or small non-protein organic molecules called coenzymes. Many of the vitamins included in a balanced diet are used by the body as coenzymes.

The shape and activity of enzymes

Enzymes [3] are very large molecules whose activity is governed by their shape. By changing their shapes it is possible to inactivate them, and thus stop certain reactions from occurring at a noticeable rate. For example, the important protein-digesting enzyme chymotrypsin occurs in an inactive form called chymotrypsinogen. Only when a few of the amino acids that make up this protein are removed does it adopt the catalytic shape of chymotrypsin. This change is triggered by the presence of food in the digestive tract. If the chymotrypsin were active all the time it would rapidly digest the intestinal wall while waiting for food to arrive.

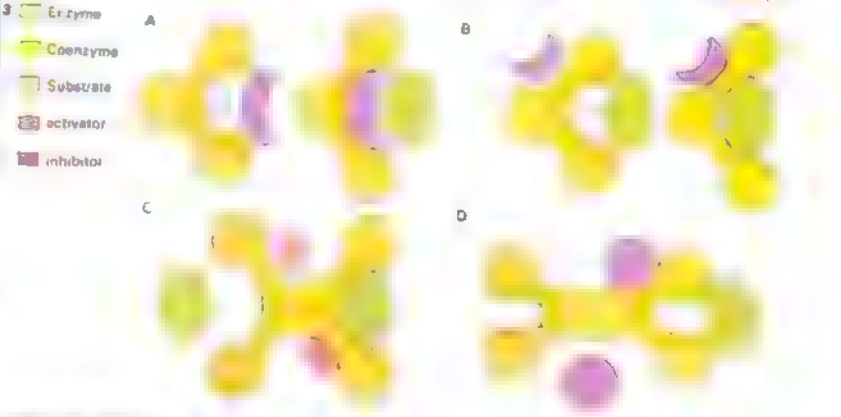
In many biochemical processes a molecule is passed from enzyme to enzyme before it becomes an end product. At each

CONNECTIONS
See also



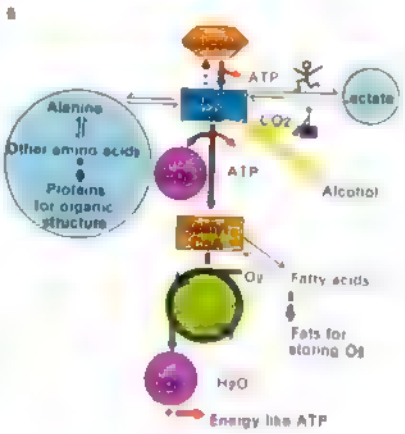
1 Adenosine triphosphate (ATP) plays a unique role in the cell, acting as a carrier between reactions that supply energy and those that use it, such as [1] cell regulation and division, [2] synthesis of important biochemicals (proteins), [3] muscle contraction, [4] transport of materials into cells, [5] conduction of nerve signals, [6] regulation of the body temperature.

2 Enzymes make molecules split up or join together much faster than they would otherwise. To work effectively, a good physical fit is needed between the enzyme and the other molecule(s) – the substrate(s). Only a part of the enzyme, the active site, often containing a coenzyme, comes into contact with other molecules. The rest of the enzyme is needed to give the active site the correct shape.



3 The shapes of enzymes can be changed by small non substrate molecules. A molecule of similar shape to a true substrate may compete for the active site [A], thereby slowing down the desired reaction. A substance at a different site [B] changes the shape of the active site affecting the fit of the substrate. Some enzymes possess more than one active site. The occupation of one or an activator may change the shape of the second site so that it can also accept the substrate [C]. An inhibitor molecule may prevent either of the sites from being used [D].

4 Enzymes can lower the "energy hump" that must be overcome in a biochemical reaction. They act as catalysts and are found in many common reactions, such as the rotting of fruit.



5 Pyruvic acid, a key biochemical, is formed during the breakdown of glucose and some amino acids. Further breakdown varies, depending on the biochemical system. In yeast, for example, the end product can be alcohol or gas to raise dough.

stage, an intermediate compound is formed. Sometimes the final product, or one of the intermediates, can combine with an enzyme farther back along the chain and switch it off. This feedback is like automation in a factory that ceases production when enough of a particular material has been made. Other small molecules may combine with an enzyme molecule to increase its activity.

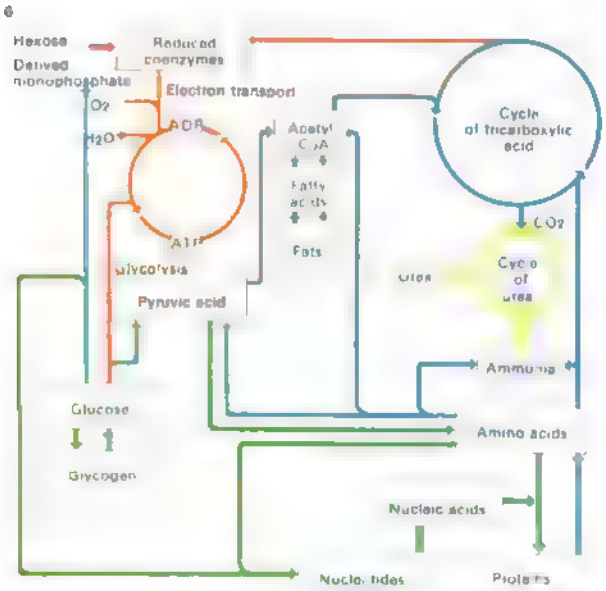
Energy from combustion

The many-step processes involving enzymes are a necessary adaptation to circumstances. Most organic chemicals are combustible. Common sugar, for example, can be burnt completely to produce carbon dioxide, water and heat. But heat energy from total combustion is of no use to living cells. To use such energy a large temperature difference is needed, as in a car engine. Living organisms have roughly the same temperature throughout – the temperature of a healthy human being rarely deviates far from 37°C (98.6°F). Consequently, chemical energy is extracted from the "combustion" process by breaking it down into a large number of small steps, to produce energy measurable (and removable) in one or two units of the cell's energy currency.

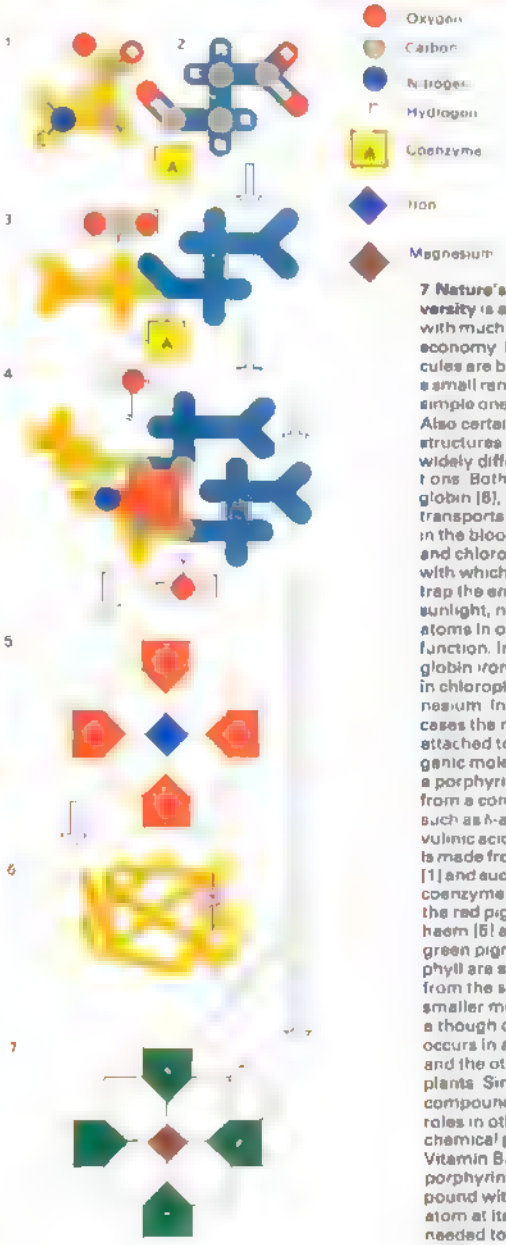
A molecule of glucose, which contains six carbon atoms, can be broken down into two three-carbon pyruvic acid molecules [5]. The process takes ten different steps, uses up two molecules of ATP and produces four new molecules of ATP. The net result of the process is therefore the production of two units of energy currency.

Pyruvic acid becomes involved in one of the key cycles of biochemistry, the citric acid (tricarboxylic acid) or Krebs cycle. It is converted into carbon dioxide and other chemicals [8]. At the same time, energy is transferred to another type of "currency" molecule, but this is soon exchanged for more ATP. Another example of energy-generation is photosynthesis, the primary process taking place in the leaves of green plants. Sunlight is absorbed by complex molecules, particularly chlorophyll, to produce "excited" molecules. These power chemical cycles that, after a number of steps, produce the ATP necessary to pay the biochemical cost of living.

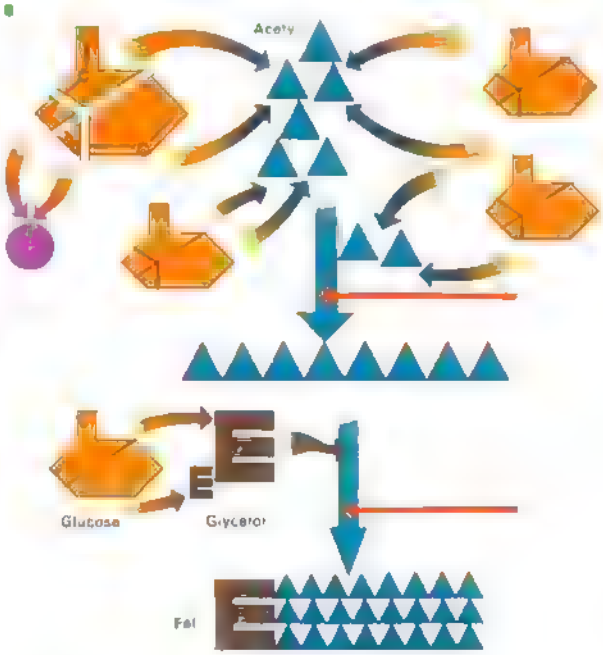
The energy that a strong man uses to lift his weights comes ultimately from the sun. Plants use light to make energy rich chemicals. Animals eat these and make even more complex substances. Human beings eat animals and plants as food, which is broken down by the digestive system into its component chemicals. These are reassembled to make various body tissues and chemicals for essential life processes. In this way these biochemicals are employed to supply both the weightlifter's muscles and the energy necessary to flex them. The detailed study of these substances and their circuitous interactions is the science of biochemistry.



6 Every cell is the site of a complex series of chemical reactions that includes both synthesis (anabolism) and breakdown (catabolism). These processes are known collectively as metabolism. Pathways that are basically catabolic are shown in blue, those basically anabolic in green. Paths directly concerned with energy production and use are in brown. Many thousands of separate chemical reactions are involved, each controlled by a different enzyme. Overall balance and control is ultimately maintained by the genetic material of the cell, which governs the production of enzymes.



7 Nature's wide diversity is achieved with much chemical economy. Large molecules are built from a small range of simple ones. Also certain key structures can have widely different functions. Both haemoglobin [6], which transports oxygen in the bloodstream, and chlorophyll [7], with which plants trap the energy of sunlight, need metal atoms in order to function. In haemoglobin iron is used in chlorophyll, magnesium. In both cases the metals are attached to an organic molecule called a porphyrin [4], made from a compound such as δ-aminolaevulinic acid [3] which is made from glycine [1] and succinyl coenzyme A [2]. Thus the red pigment haem [5] and the green pigment chlorophyll are synthesized from the same smaller molecules, although one of them occurs in animals and the other in plants. Similar compounds play key roles in other biochemical processes. Vitamin B_{12}, a porphyrin-like compound with a cobalt atom at its centre, is needed to prevent pernicious anaemia.



8 All cycles interlink so that the energy from breaking down one type of chemical can be used to make another. If a man overeats he will put on excess weight – in other words he becomes fat. Yet as the dieter is frequently warned, it is carbohydrates such as sugar, more than fats, that increase weight. Pyruvic acid is the key to this apparent paradox. When it is broken down, in addition to carbon dioxide, a substance called acetyl coenzyme A can be formed. The acetyl part (the triangles) is the basic building block for the fatty acids which, together with glycerol (E-shape), make up oils and fats such as the one pictured at the bottom.

Polymers: giant molecules

The metals that make up a car's body and engine are chemically quite different from the oil products that power and lubricate them. Nature is more economical; the same few elements used to build living organisms are also those that trap and transport the energy, all originating from sunlight

Proteins: polymers of amino acids

The important structural molecules of plants and animals are polymers, very large molecules known as macromolecules, made by joining together a succession of simple chemical building blocks. Proteins, for example, are polymers of amino acids which are small molecules, each containing an amino group and a carboxylic acid group These two groups can react with one another to form a chemical bond. As a result, different amino acids can be linked through these groups in very large numbers. A small protein, such as insulin [1], may be made up of only 50 or so amino acids, but on the other hand many proteins contain hundreds of individual amino-acid units

Animals employ proteins [2] both to build tissues and in the biochemical processes which take place in them. Collagen, for example, is a common structural protein One of its jobs is to provide materials for tendons, which are essential to movement A tendon "rope" of intertwined collagen molecules can have the strength of light steel wire. Another structural protein, keratin [8] occurs in hoof, hair, horn and feathers, and actin and myosin are important constituents of muscle Proteins also supply the major (in some cases the sole) component of enzymes, the cell's catalysts which speed up specific biochemical reactions, and the antibodies which fight infective micro-organisms

Essential to these differing roles are the various physical structures of proteins. When amino acids link together, they do not just form long chains. According to the shapes and the chemical properties of the side-chains of the individual amino acids used in a protein's make-up, it can be long and thin, or compact and globular The structures contain electrically charged groups and, in addition sulphur atoms can form bridges between amino acids. In insulin, sulphur atoms bridge adjacent chains of amino acids. In cytochrome C, a sulphur atom attaches a non protein organic molecule (in this case haem) to a protein molecule

Polysaccharides: polymers of sugars

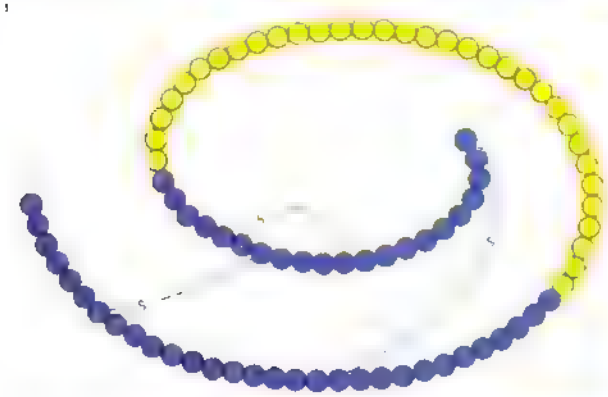
The essential structural components of plants are polysaccharides, polymers of the small sugar molecules which provide most of the energy for cells. Not surprisingly, some polysaccharides are also used as a convenient means of storing energy

It has been estimated that up to 50 per cent of the carbon atoms incorporated in plant tissues are in molecules of cellulose, a structural glucose polymer. In some forms cellulose has commercial value; cotton, for example, is 98 per cent cellulose

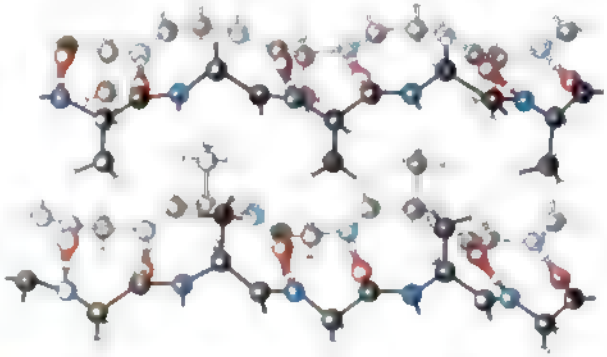
The main energy reserve polymer of plants is starch, which is composed of two polymers – amylose and amylopectin Like cellulose, amylose is a straight-chain polymer of glucose. The only difference between the two molecules is in the shape of the chemical bond that links the units together. This single difference is enough to make starch a readily

CONNECTIONS

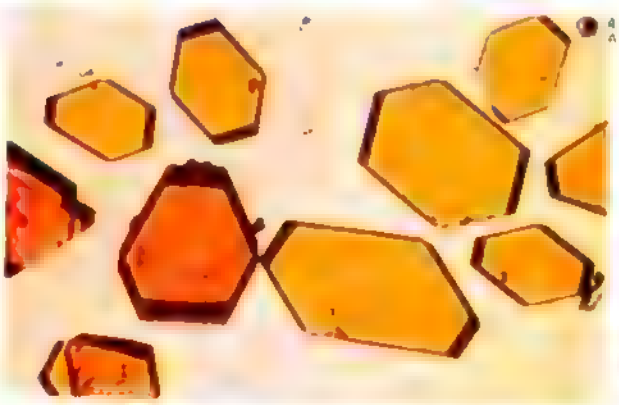
See also



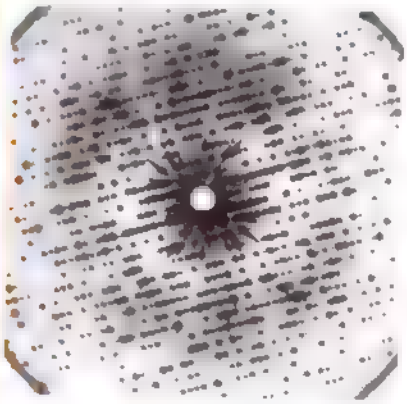
1 Insulin is formed by the removal of 33 amino acid units (yellow) from the protein proinsulin. The two chains of the insulin molecule are bound covalently by sulphur-sulphur bridges

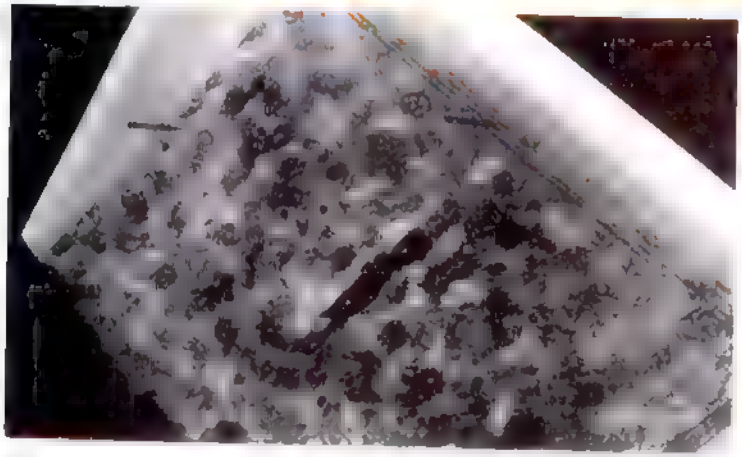


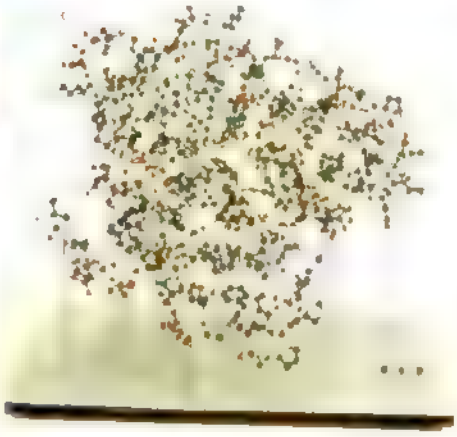
2 In all proteins, amino acids made of carbon (black), nitrogen (blue), hydrogen (white) and oxygen (red) join to form chains Hydrogen bonds (red) can hold chains together – as in this pleated β-sheet structure of silk



3 Some proteins can be purified sufficiently to form crystals Once the crystals have grown large enough their molecular structure can be determined by X ray crystallography Myoglobin, a protein which is involved in the carrying of oxygen in muscles, occurs in many species including sperm whales from which the crystals illustrated here (enlarged 40 times) were obtained







4 The diffraction pattern produced when a beam of X-rays is diffracted (scattered in a periodic manner) by a single crystal can be detected on photographic film [A] The position and intensity of each spot is measured Thousands of diffraction spots, from several films, may have to be analysed with the aid of a computer This provides information from which electron density maps can be drawn Made from sheets of perspex stacked on top of one another, these maps show where particular atoms in the protein molecules are located [B] It is then possible to construct an accurate 3-dimensional model [C] From the structure of myoglobin much can be learnt about its function The white tube traces the path of the "Backbone chain" of the amino acid polymer It folds around a raft-like "heme-group" containing an iron atom (large red ball) at its centre This binds and releases oxygen in response to its environment

digestible dietary ingredient and cellulose completely indigestible to human beings. In amylopectin, there are chemical links at more than one point on some of the glucose units, so that a branched-chain polymer is produced. The same sort of structure occurs in glycogen (animal starch), the glucose polymer used by animals for energy storage.

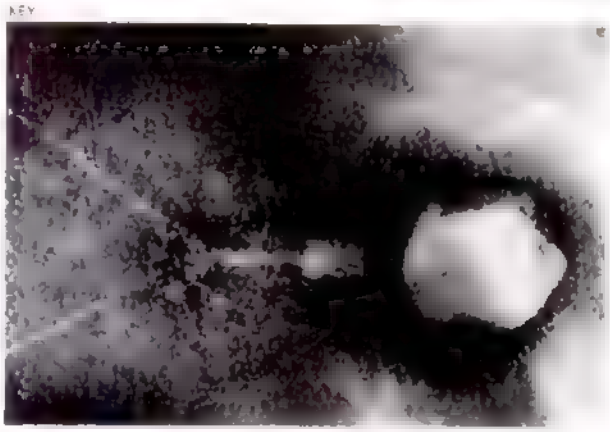
Many other sugars, apart from glucose, can form polysaccharides. Chitin, the hard shell material of insects, crabs and lobsters, is a polymeric aminosugar. Alginates, important food additives which keep the head on beer and give dehydrated soups their thickness, are polysaccharides from seaweed, while the pectins, which are widely used in jam-making, occur notably in apples. Natural adhesives, such as gum arabic, are polysaccharides, as is heparin, an important substance that prevents blood clots; it is often used in the treatment of thrombosis.

Nucleic acids

Although not present in such large quantities in most cells as proteins and polysaccharides, the most important macromolecules are the nucleic acids [7]. These make up the genetic material which controls each cell, making it not only a man-cell or a mouse-cell, but a man-liver-cell or a mouse-tail-cell. Nucleic acid polymers are able to reproduce themselves accurately, therefore allowing any species to produce more of its own kind. They also control the chemical building of proteins. As the latter effect includes production of various enzymes, nucleic acids control all other chemical building up and breaking down in living tissues.

Nucleic acid polymers [7] can be thousands of units long. The basic repeating unit is made up from a nitrogen-containing base and a phosphate group. Both of these are attached to one of two types of sugar: ribose (in RNA) or deoxyribose (in DNA). Because of the chemical properties of the bases, particularly those in DNA, it is possible for two strands of nucleic acids to fit together readily to form the "double helix".

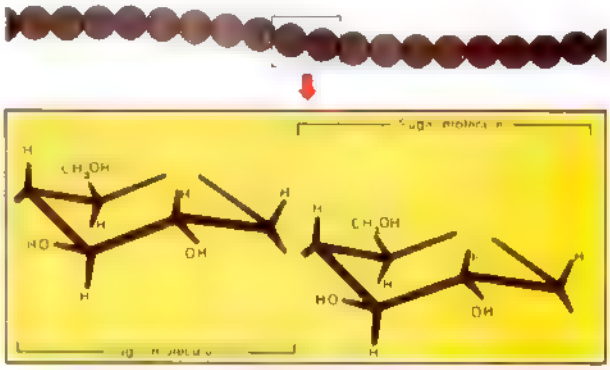
Synthetic rubber and plastics are also polymers – often man-made copies of natural molecules. Glass and similar substances are inorganic polymers.

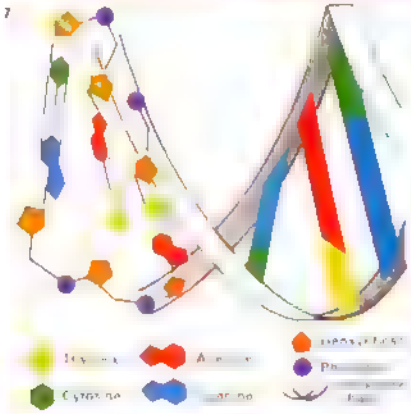


On the margin of life lie the viruses, each made up of a few macromolecules, all of which can be defined in purely chemical terms. Yet when placed in a living cell, a virus is able to take over that cell's biochemical machinery and make it reproduce virus components. This electron micrograph of a bacteriophage – a virus that attacks bacteria – shows the shape of its protein molecules. Inside the diamond-shaped head is the bacteriophage's nucleic acid which directs the build-up of further examples of both itself and the protein components after it has infected a cell.

5 Polysaccharides, proteins and polynucleotides are the basic polymers of life. They are not, however, the only ones. Natural rubber, for example, is a polymer made up mainly of repeating units of the unsaturated hydrocarbon isoprene. From 1,000 to 5,000 such units join together in a single molecule of rubber. The cell walls of some bacteria are made from a combination of sugar and amino acid molecules, to form a mixed polymer. This photograph, taken at a magnification of 280,000 times with the aid of an electron microscope, shows part of the outer cell wall of the bacterium *Clostridium thermohydrosulfuricum*. At this magnification it is possible to see the individual subunits, arranged in regular rows, that make up the surface of the bacterium's cellular wall.

6

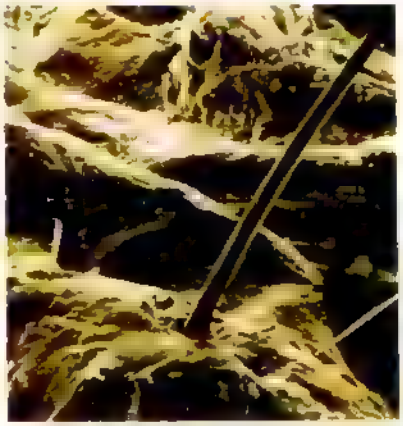




6 Monosaccharides, the simplest of the sugar molecules, can join together to form very large molecules such as starch. According to where the links form, the macromolecule may be a single chain (as shown) or branched.

7 Deoxyribonucleic acid (DNA) is the master molecule of life. It occurs as a double helix in which two complementary strands of polymer are held together by hydrogen bonding. This bonding occurs between the nitrogen containing bases which form part of the nucleic acid unit.

8



8 Hair is composed mainly of the protein keratin, which also occurs in feathers and skin. Although these hairs are magnified many times, it is still not enough to make individual keratin molecules visible. Whether a person's hair is straight or curly depends on the tertiary structure of the keratin molecules.

Scientific terms and tables

CHEMICAL ELEMENTS

[illegible]

Element	Symbol	Atomic number	Atomic weight	Melting point (°C)	Boiling point (°C)	Relative density or density
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	85	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
chromium	Cr	24	51.996	1890	2482	7.19
cobalt	Co	27	58.9332	1495	2870	8.9
columbium	Cb		see niobium			
copper	Cu	29	63.546	1083	2595	8.96
curium	Cm	96	*247	1340		13.51 [illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Fm	100	*257			
[illegible]	F	9	18.9984	219.62	188.14	1.696 kg m⁻³
[illegible]	Fr	87	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2660	19.30
[illegible]	Hf	72	178.49	2150	5400	13.31
[illegible]	He	2	4.0026	272.2	268.6	0.178 kg m⁻³
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	H	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	184.35	[illegible]
iridium	Ir	77	192.22	2410	4130	22.42
iron	Fe	26	55.847	1539	2800	7.90
krypton	Kr	36	83.80	156.6	152	3.733 kg m⁻³
lanthanum	La	57	138.9055	920	3454	6.17
lawrencium	Lr	103	*267			
lead	Pb	82	207.19	327.3	1750	11.3
lithium	Li	3	6.941	179	1317	0.534
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Mg	[illegible]	[illegible]	[illegible]	[illegible]	1.738
[illegible]	Mn	[illegible]	[illegible]	1244	[illegible]	[illegible]
[illegible]	Md	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Hg	[illegible]	[illegible]	[illegible]	356.58	[illegible]
[illegible]	Mo	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Nd	60	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Ne	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Np	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Ni	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Nb	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	N	[illegible]	[illegible]	209.86	[illegible]	[illegible]
[illegible]	No	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Os	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	O	8	[illegible]	[illegible]	182.96	[illegible]
[illegible]	Pd	46	[illegible]	[illegible]	[illegible]	[illegible]

CHEMICAL ELEMENTS

Element	Symbol	Atomic number	Atomic weight	Melting point (°C)	Boiling point (°C)	Relative density or density
[illegible]	P	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Pt	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Pu	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Po	84	[illegible]	254	[illegible]	[illegible]
[illegible]	K	19	[illegible]	[illegible]	[illegible]	0.86
[illegible]	Pr	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
promethium	Pm	61	*147	1080	2460	
[illegible]	Pa	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Ra	88	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Rn	86	[illegible]	[illegible]	61.8	[illegible]
[illegible]	Re	75	[illegible]	3180	[illegible]	[illegible]
[illegible]	Rh	45	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Rb	37	[illegible]	38.89	688	[illegible]
[illegible]	Ru	44	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Sm	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Sc	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Se	[illegible]	[illegible]	[illegible] grey	684.9 grey	[illegible] grey
[illegible]	Si	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Ag	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Na	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Sr	38	[illegible]	[illegible]	[illegible]	[illegible]
sulphur	S	16	32.064	112.8 (rhombic) 119.0 (monoclinic)	444.6	[illegible]
tantalum	Ta	73	180.9479	2996	[illegible]	[illegible]
[illegible]	Tc	[illegible]	[illegible]	[illegible]	5030	[illegible]
[illegible]	Te	[illegible]	[illegible]	449.5	989.8	[illegible]
[illegible]	Tb	[illegible]	158.9254	[illegible]	[illegible]	[illegible]
[illegible]	Tl	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
thorium	Th	90	232.0381	1750	3800	[illegible]
thulium	Tm	69	168.9342	1645	1727	9.31
tin	Sn	50	[illegible]	[illegible]	[illegible]	[illegible] grey [illegible] white
[illegible]	Ti	[illegible]	47.90	[illegible]	[illegible]	4.54
[illegible]	W	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	U	92	[illegible]	[illegible]	[illegible]	18.95
vanadium	V	23	50.9414	1890	3380	[illegible]
wolfram	W	74	see tungsten			
[illegible]	Xe	[illegible]	[illegible]	[illegible]	[illegible]	5.88 [illegible]
[illegible]	Yb	[illegible]	173.04	824	[illegible]	6.97
[illegible]	Y	[illegible]	88.9059	[illegible]	[illegible]	[illegible]
[illegible]	Zn	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Zr	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(Element 104)		104				
(Element 105)		105				

BEAUFORT SCALE

			Speed (knots)
[illegible]	[illegible]	[illegible]	
[illegible]	[illegible]	[illegible]	1
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	Fresh breeze	[illegible]	16
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
8	[illegible]	[illegible]	28 33
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

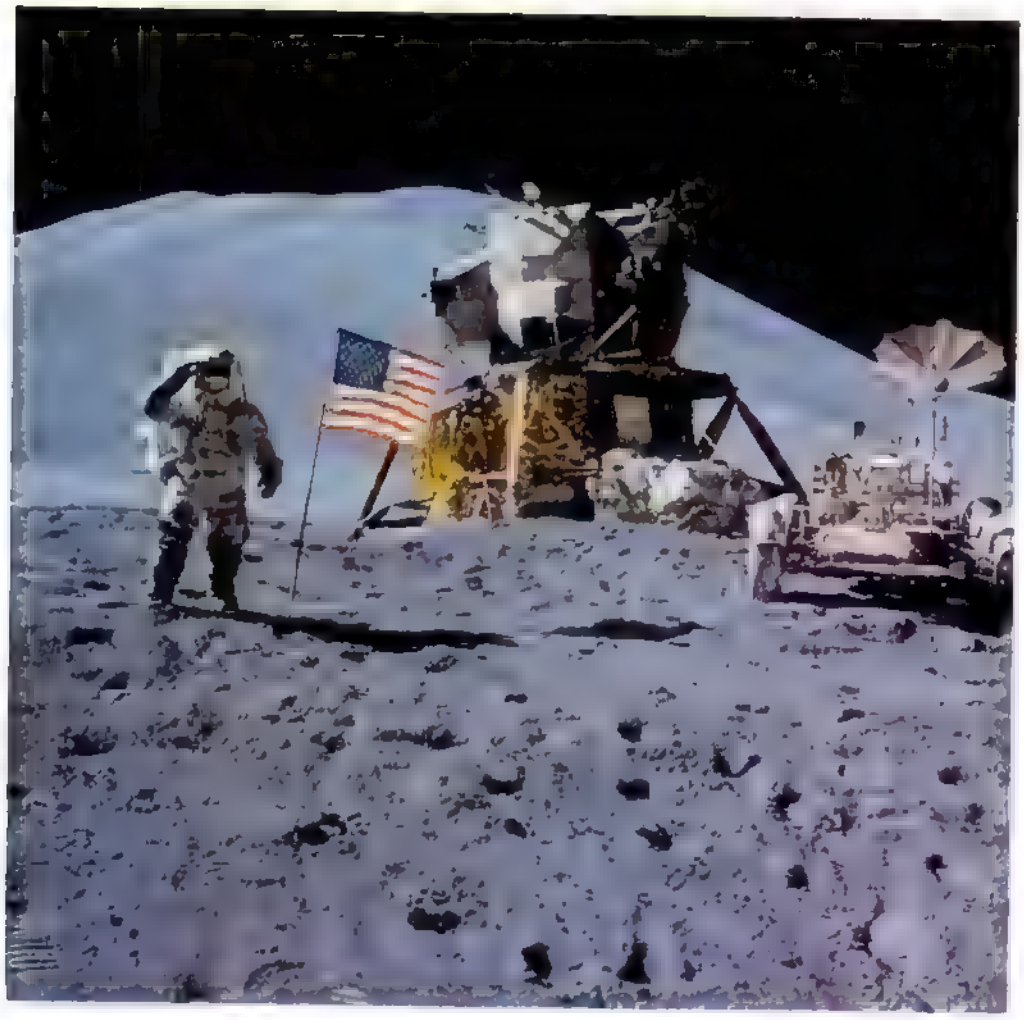
ALLOYS

Name	Typical composition	Properties	Uses
Brass	70% Cu 30% Zn	malleable, corrosion resistant, good conductivity	rust free parts, electrical contacts etc
Bronze	90% Cu 10% Sn	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
Stainless steel	85.1% Fe 13.7% Cr 0.3% C + Ni, Mn, Si	high strength and corrosion resistance	cutlery, chemical plant, ball bearings

THERMOMETER COMPARISONS (CELSIUS/FAHRENHEIT)

C	F	C	F	C	F
17.8	0	21.1	70	48.9	120
6.7	20	25	77	50	122
1.1	30	26.7	80	54.4	130
0	32	30	86	55	131
4.4	40	32.2	90	60	140
5	41	35	95	65	149
10	50	37.8	100	65.6	150
15	59	40	104	70	158
15.6	60	43.3	110	71.1	160
20	68	45	113	75	167

5 Space Sciences



The restless sky

Astronomy is the oldest of all the sciences. It was a natural and long-held assumption that the Earth must be flat and lie at rest in the centre of the universe, with the entire sky revolving round it once every 24 hours, but at an early stage it was also clear that many celestial bodies had their own relative motions across the heavens.

Movements in our sky

The Moon was seen to shift quickly in position against the starry background and the Sun, of course, had its own motion. Then there were also the occasional spectacular phenomena, sometimes the Sun was blotted out during a solar eclipse and sometimes the Moon became strangely dim when full. It was not then known that a solar eclipse occurs when the Moon passes between the Sun and the Earth, casting a shadow on the Earth, and a lunar eclipse occurs when the Sun, Earth and Moon are in line and the Moon enters the Earth's shadow; but it has been suggested that some of the old stone circles were primitive eclipse computers [Key].

The Greeks recognized that the five bright planets – Mercury, Venus, Mars, Jupiter and Saturn – moved against the stars and were thus fundamentally different from them. The constellation patterns remained unchanged over long periods and it was originally thought that the stars were fixed onto a crystal sphere revolving round the Earth.

The planets were regarded as being closer to the Earth and it was thought that they, like the Sun and Moon, moved round the Earth between the Earth's surface and the sphere of the so-called fixed stars. The old system was perfected by Ptolemy (c. AD 90–168). In the Ptolemaic system [1], all celestial orbits were assumed to be perfectly circular; but as the observed movements of the planets did not conform to the idea of circular motion at uniform velocity, it was necessary to introduce complications – the epicycle that described the movement of a planet is a small circle the centre of which (the deferent) itself moved round the Earth in a circle.

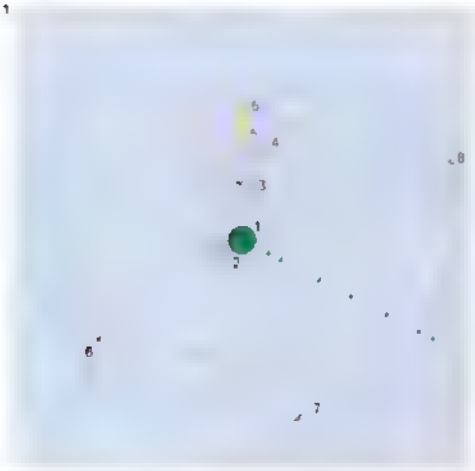
A few of the earlier Greek philosophers, notably Aristarchus (310–230 BC), had believed that the Earth moved round the Sun, but the Sun-centred or heliocentric theory was generally rejected until the work of Copernicus (1473–1543), a Polish canon, made its impact in the sixteenth century. Copernicus took the drastic step of removing the Earth from its proud central position and put the Sun there instead [2]. However he retained perfectly circular orbits and was even compelled to retain epicycles. The modern phase of astronomy dates from the publication of his great book, *De Revolutionibus Orbium Caelestium*, in 1543.

Revolutionary outlooks

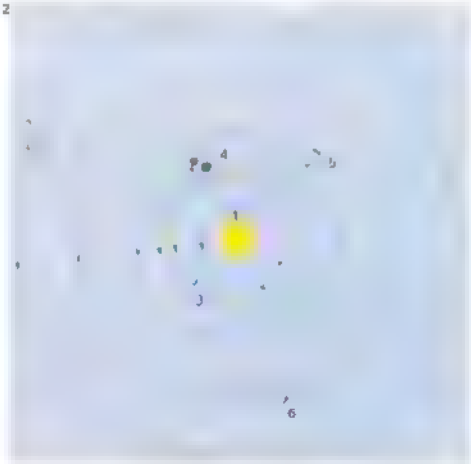
Inevitably, the Copernican system was strongly opposed. Tycho Brahe (1546–1601), the Danish astronomer who was the most thorough observer of pretelescopic times, believed the planets moved round the Sun while the Sun and Moon moved round the Earth [3]. When Tycho died his observations of the positions of the stars and the movements of the planets came into the possession of his last assistant, Johannes Kepler (1571–1630). After years of work, Kepler realized that the planets move round the Sun not in circles but in ellipses, and be-

CONNECTIONS

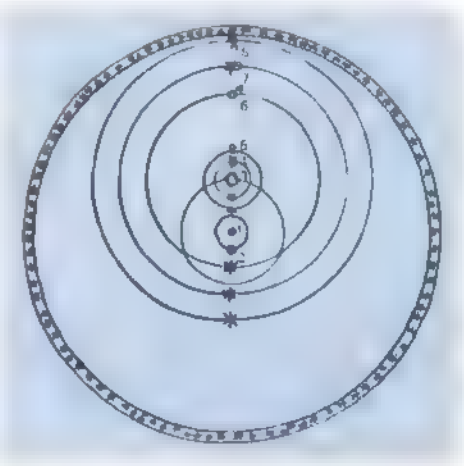
See also
Measuring the
Mathematics and



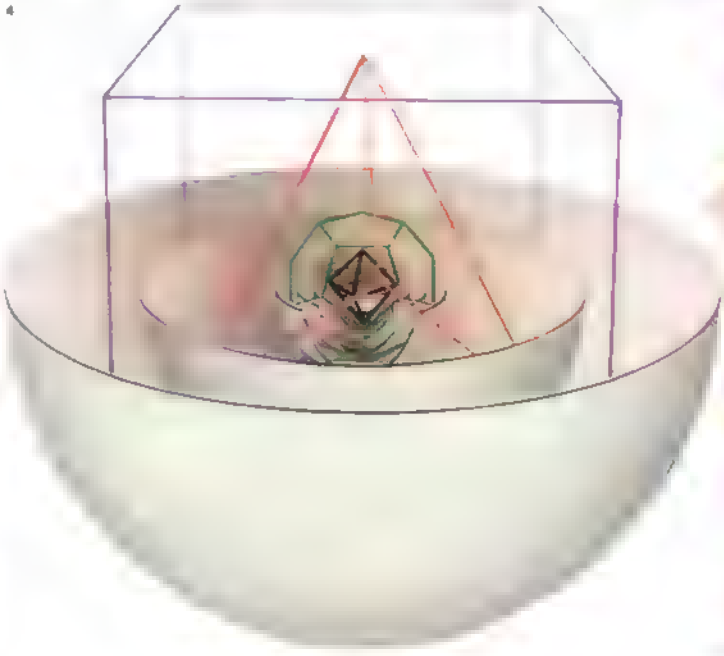
1 In the Ptolemaic system, the Earth [1] lies at rest in the centre of the universe. Round it move the Moon [2], Mercury [3], Venus [4], the Sun [5], Mars [6], Jupiter [7] and Saturn [8], each body moving in a small epicycle.



2 The Copernican theory places the Sun [1] in the centre of the Solar System orbited by Mercury [2], Venus [3], the Earth [4], Mars [5], Jupiter [6], and Saturn [7]. Copernicus's book was published in 1543. His theory met strong opposition from the Church, and religious persecution persisted for a century. Copernicus retained both circular orbits and epicycles.



3 Tycho Brahe believed that the Earth [1] was the centre of the Solar System, orbited by the Moon [2] and the Sun [3]. The planets – Mercury [4], Venus [5], Mars [6], Jupiter [7] and Saturn [8] – moved round the Sun, with the stars beyond.



4 Kepler's theory of "five regular solids" shows that his ideas provided a link between the past and the present. He believed that the five regular solids – the cube [A], tetrahedron [B], dodecahedron [C], icosahedron [D] and octohedron [E] – could be fitted inside the orbits of the various planets. He reasoned that there were only five such solids and exactly five spaces between the six planets known at that time: Mercury, Venus, Earth, Mars, Saturn and Jupiter. It was his brilliant work, based upon observations made by Brahe, that showed that the Sun, not the Earth, was the centre of the Solar System. Kepler was both a mathematical genius and an astrological mystic.

5 According to Newton, were it not for the Earth, the Moon would move in a given period from A to A1, but because of Earth's pull, the actual movement is from A to A2. The Moon keeps "falling" towards the Earth, although it comes no closer to us. This illustrates the law that a body will continue in a state of rest, or uniform motion in a straight line, unless acted upon by an outside force. This law was laid down in Newton's *Philosophiae naturalis principia mathematica* (1687).

tween 1609 and 1618 he published his three fundamental laws of planetary motion. The first law states that the orbit of a planet is an ellipse, with the Sun at one of the foci. The second law states that a planet moves at its fastest when it is closest to the Sun and the third law provides a definite relationship between a planet's sidereal period (that is to say, the time taken for the planet to complete one journey round the Sun) and its distance from the Sun. Using Kepler's laws it became possible to draw up a scale map of the Solar System. When one distance could be determined absolutely, the distances of all the rest could be obtained by calculation.

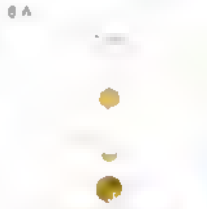
The revolution in outlook was completed by Isaac Newton (1642–1727), whose book – usually called the *Principia* [5] published in 1687 – laid the foundations of all subsequent work. By then the distance of the Sun from the Earth was known with reasonable accuracy and it had become clear that the Solar System was a very small part of the universe as a whole. The stars were known to be suns in their own right and to be so far away that their apparent individual or proper motions were to an observer very slight indeed.

Edmond Halley (1656–1742), friend and contemporary of Newton, used ancient observations to demonstrate that a few of the bright stars had shown relative shifts over the centuries, so that even the constellation patterns could change gradually with time.

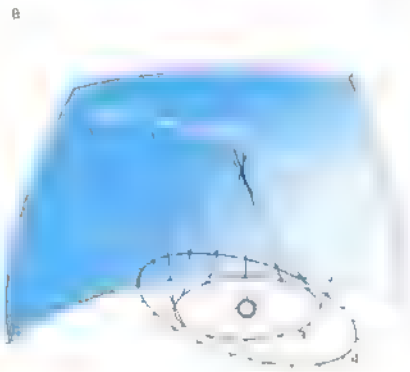
The scale of the universe

The scale of the universe was established only much later, when distances to stars began to be measured. In 1838 Friedrich Bessel (1784–1846) first measured such a distance to a nearby star (in Cygnus) and found it to be about 96 million million kilometres (60 million million miles) away. Since light takes about 11 years to cover this distance, the star is said to be 11 light-years away. Most stars are much more remote than this, but modern techniques enable astronomers to measure their proper motions from year to year. The old name of "fixed stars" is misleading; all the stars are moving at high velocities relative to each other. In our own century it has been shown that our Galaxy is itself one of many; there are millions of other galaxies.

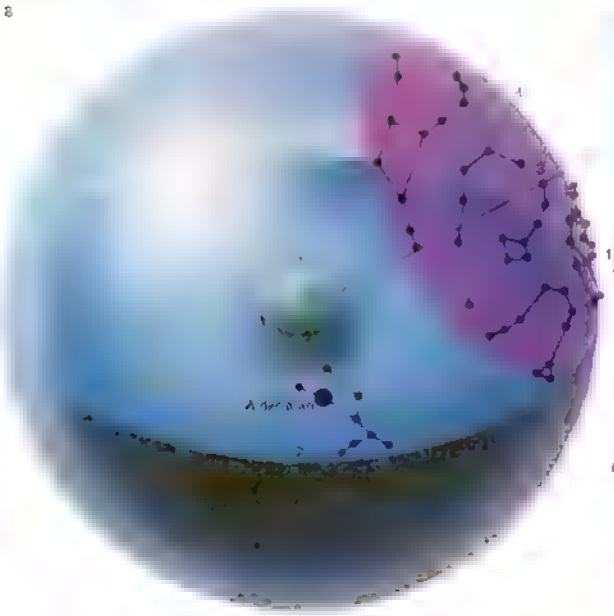
Stonehenge (A), the famous megalithic "stone circle" in Wiltshire, England, is made of standing stones with alignments that may have astronomical significance. Many of the alignments among the concentric rings (B) point to spots on the horizon where the Moon and Sun rise and set. For example stone 1 as viewed from stone 2 marks the point where the Moon sets in its most northerly position in midwinter. Viewed from stone 3, stone 1 marks the midsummer sunrise. Stonehenge may have been an early primitive computer predicting eclipses for religious ends or, more practically, fixing the solstices, which were important in the agricultural calendar.



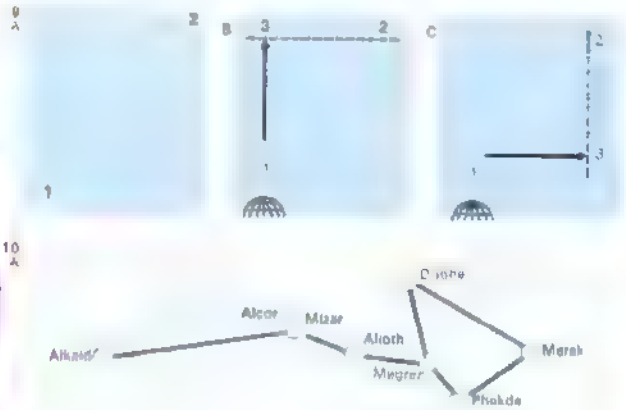
6 The two planets whose orbits lie within that of the Earth – Mercury and Venus – show lunar-type phases and remain in the same area of the sky as the Sun (A). An inner planet is at inferior conjunction (1), its dark side is turned towards Earth so it appears "new". When on the far side of the Sun (2) it is full. The synodical period, or mean interval between successive inferior conjunctions, is 116.9 days for Mercury and 583.9 days for Venus. The orbits of the Earth (4) and Venus (3) are shown (B) and the white line indicates the apparent motion of Venus in the sky. Mercury behaves in a similar way.



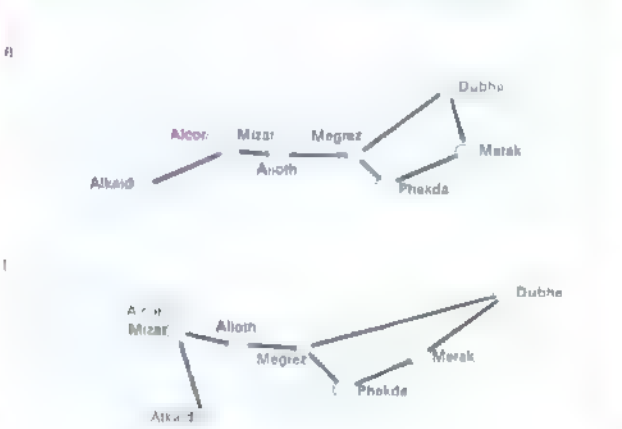
7 Orbits of superior planets (A), beyond Earth's orbit, reach opposition (1) and conjunction (2). Apparent motion (B) of a superior planet (4) in relation to Earth (3) appears temporarily retrograde.



8 Astronomers visualize the stars, for the sake of convenience, as lying on the inside of a sphere centred on the Earth. To an observer on Earth (1), the horizon becomes a circle (2) when projected on to this celestial sphere. As the Earth rotates west to east, the sky seems to move east to west, taking the stars, such as Aldebaran in Taurus, with it. The north pole of the sky (3), which is indicated approximately by Polaris, is stationary to an observer in the Earth's Northern Hemisphere. Stars in the mauve area (4), drawn for an observer at latitude N50°, remain permanently above the horizon and are called circumpolar stars. The circumpolar area depends on the observer's latitude.



9 The real movement of a star in space includes actual motion (A) where the star moves from 1 to 2 in a given period. Radial motion (B) is when the star moves from 1 to 3 if receding (positive radial motion), or 3 to 1 if approaching (negative radial motion). Proper motion (C) is the term used for the transverse movement (1 to 3) against the background of more distant stars. (A) combines (B) and (C).



10 To the naked eye constellations appear unchanged over thousands of years, but over a sufficiently long period the proper motions show up. The seven main stars of Ursa Major, including the double star Mizar, are shown as they were 100 000 years ago (A), as they are today (B), and as they will appear 100,000 years hence (C). It is evident that Dubhe and Alkaid are moving in an almost opposite direction to that of the other five.

Measuring the restless sky

The size of the universe is almost unimaginable. It is easy to comprehend the distance from London to New York, or from New York to Australia, and the Moon does not seem impossibly remote, because its distance is only ten times greater than that of a journey right round the Earth. But any attempt to visualize what is meant by "a million kilometres" is doomed to failure – and a million kilometres is a very short distance on the cosmic scale

Early estimates of distance

The ancients had no idea of scale (it was once thought that the diameter of the Sun was only 70cm [27in]), but they were able to measure the size of the Earth itself with remarkable accuracy [1] As soon as the old concept of an Earth-centred universe was abandoned, distance estimates became much more realistic Giovanni Cassini (1625–1712) gave the distance between the Earth and the Sun as 138 million kilometres (86 million miles), which was approaching the true figure Astronomers then decided that the Earth–Sun distance was to be *the* astronomical unit, and measuring it became a major task [Key]

The basis of any method of computing this distance was Kepler's third law, which established a definite relationship between the revolution period of a planet and its distance from the Sun. The revolution period of the Earth was known to be 365 25 days and the periods of the other planets could be found from observation – 687 days for Mars, and so on – with the result that a complete scale model of the Solar System could be drawn up Thus if it were possible to obtain the distance from Earth to any planet (eg Mars or Venus), Kepler's third law would give the Earth–Sun distance [4]

The parallax principle

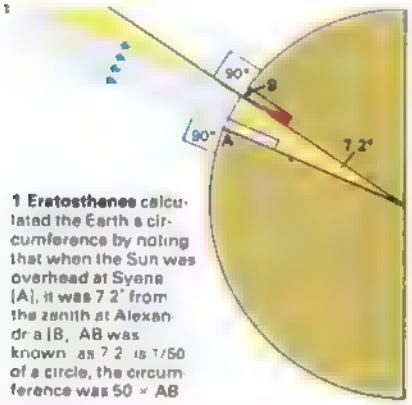
The obvious way to calculate the distance from the Earth to one of the planets was to use parallax [2], a method also used by surveyors. If a not-too-distant object is observed against a background of more remote objects, its position will seem to alter according to the position of the observer If the distance between the two observation points is known, and the respective angles formed by the object with the line are measured, an astronomer can calculate by trigonometry the height of the triangle thus formed, ie the distance of the heavenly body

Edmond Halley (1656–1742), the second Astronomer Royal, proposed to make use of transits of Venus – the rare occasions when Venus passes in front of the Sun as seen from Earth, and appears as a black spot against the solar disc – to determine the planet's absolute distance. Attempts made during the transits of 1761 and 1769 were only partly successful (Captain Cook's voyage to the South Seas in the latter year was for the express purpose of observing the transit) The next transits were those of 1874 and 1882, all measurements agreed in making the astronomical unit about 150 million kilometres (93 million miles)

When Venus passed across the face of the Sun, however, it seemed to draw a strip of blackness after it, distorting its shape, when this so-called "Black Drop" vanished, the transit had already begun, which meant that the measurements were subject to considerable error. In 1877 efforts were made to determine the parallaxes of three asteroids –

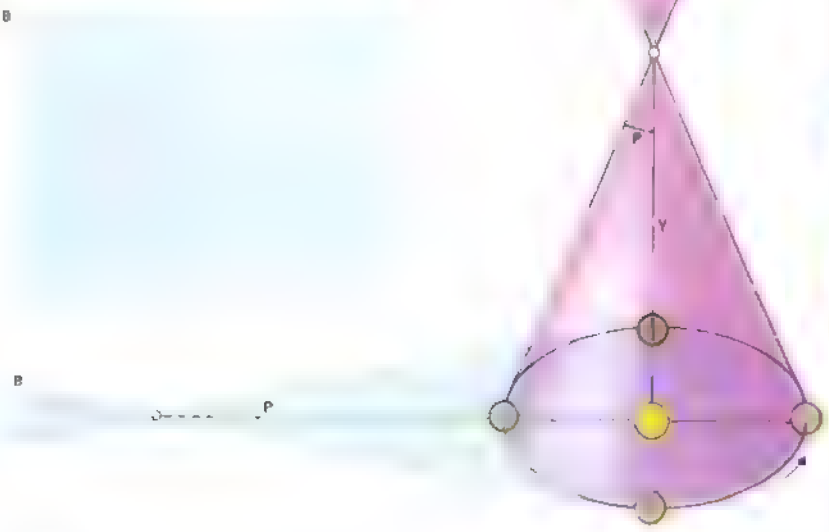
CONNECTIONS

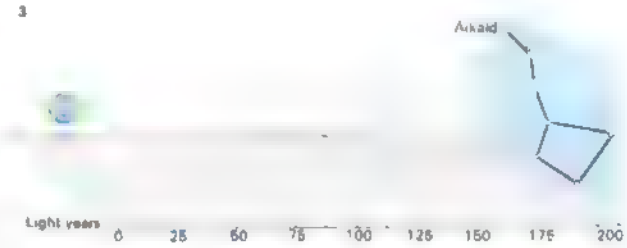
See also



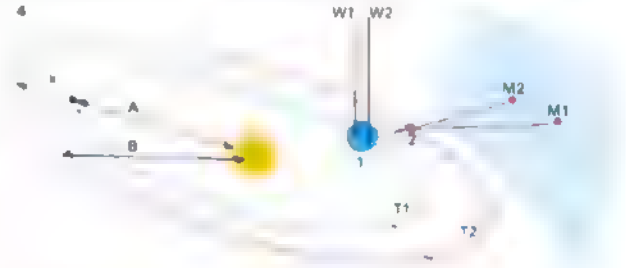
1 Eratosthenes calculated the Earth's circumference by noting that when the Sun was overhead at Syene (A), it was 7 2° from the zenith at Alexandria (B, AB was known as 7 2 is 1/50 of a circle, the circumference was 50 × AB

2 Apparent parallaxes of stars are illustrated in these diagrams. The parallaxes are measured for the apparent movements of relatively nearby stars against a background of more remote objects, its position will seem to alter according to the position of the observer If the distance between the two observation points is known, and the respective angles the parallactic motion will take the form of a to-and-fro straight line P is the angle of parallax and from this the distance of the star can be measured. The main difficulty is, of course, that the angle P is always very small With modern photography and other techniques we can now "see" out to at least 200 light years, but at greater distances the parallax shifts are swamped by observational errors.





3 Stars in Ursa Major seem to the casual observer to be the same distance from the Earth, there is no "three dimensional" effect In fact the stars in any particular constellation are not necessarily associated The diagram shows the relative distances of the seven main stars in Ursa Major Alkaid (210 light years) is easily the most remote Mizar lying next to it in the sky is only 88 light-years away



4 A planet's distance from Earth (x) can be calculated using triangulation. A and B are the distances of Earth [1] and the planet [2] from the Sun. M1 and M2 are the planet's apparent positions as seen from W1 and W2 on Earth The ratio of the angles at M1 and M2 is the same as the ratio of the distances W1–W2 and x. Using Kepler's third law $(T1/T2)^2 = (A/B)^3$, where T is the period of orbit, A and B can also be found

5 The distance of Venus can be measured by radar Once this is known Kepler's third law can be used to find the distance of the Earth from the Sun. Radar pulses of differing wavelengths are reflected from Venus to the receivers on Earth. The time-lag between transmitting and receiving echoes gives the distance travelled, provided allowances have been made for the delaying of the echoes by the ionosphere

Iris, Victoria and Sappho – which appeared as star-like points. In 1931, when the asteroid Eros passed within 24 million kilometres (15 million miles) of the Earth, there was a worldwide programme to determine its parallax precisely

New methods involve the use of radar Radar transmits a pulse of energy to a remote object and receives an echo from it: radio waves travel with the speed of light and, because this speed is constant, the time-lapse between the transmission and the arrival of the echo enables the distance of the object to be calculated. The planet Venus is contactable by radar [5] This new method gives the length of the astronomical unit as 149,600,000km (92,957,000 miles)

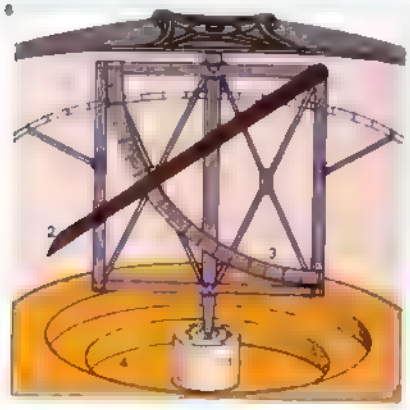
The problem of mapping the stars

Star distance presented different problems and here again parallax was used If a nearby star is observed over a six month interval it will show a parallax shift against the background stars, because during the interim the Earth will have moved from one side of its orbit to the other, giving a "baseline" of 300 million kilometres (186 million miles). This method was first applied in 1838, when Friedrich Bessel (1784–1846) showed that a faint star in Cygnus lay at a distance of 11 light-years. (One light-year is the distance travelled by light in one year; that is, 9,460,000 million kilometres [5,880,000 million miles])

The parallax methods work well for the nearer stars, but beyond a few hundreds of light years the shifts become too fine to be measured and less direct methods must be used. Spectroscopic work will give the real luminosity of a star and this, compared with its apparent brightness (or visual magnitude), can give the distance It is now known that the diameter of our Galaxy is about 100,000 light-years

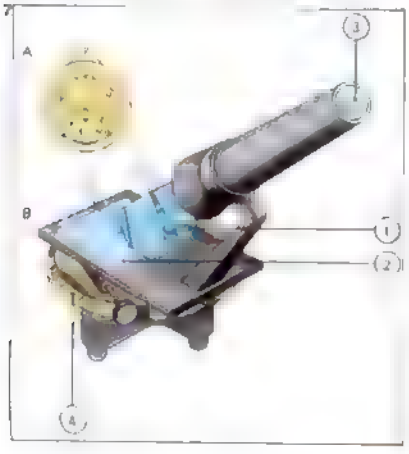
But our Galaxy is not the only one The hazy patches known as nebulae are of two kinds: some can be resolved into stars, others cannot In 1845 Lord Rosse (1800–67), using his 72in (183cm) telescope, found that many of the starry nebulae are spirals and it has now been found that spirals are external systems, millions of light-years away

Aristarchus, the Greek astronomer of the 3rd century BC is said to have been the first man to propose a heliocentric theory of the universe He was also able to measure the relative distances of the Sun and Moon When the Moon is at first quarter [1] the angle it makes with the Sun [3] is near 90 By measuring the angle at the Earth [2] Aristarchus could determine from the triangle the relative distances from Earth He found the angle to be 87°, instead of the true value of 89°52' but a small error can lead to a large discrepancy in the ratio of the distances of the Sun and Moon respectively from the Earth Aristarchus ratio was 19 to 1, the true ratio is 390 to 1



6 Tycho Brahe's quadrant was one of the instruments used for measuring star positions. This quadrant, used between 1576 and 1598 was mounted on a central pillar [1]; a pointer [2] with sights rotated against a graduated metal circle 3] The well [4] accommodated the observer at various levels determined by the position of the pointer Modern work depends on fundamentals like those established by Tycho

7 Ancient astrolabes [A] had simple pointers and scales to measure the altitudes of stars and other objects in the sky In the modern astrolabe [B] the light strikes a prism [1] and a mercury surface [2] forming a double image along the collimation line [3] The images separate when the object moves and this separation is measured by the azimuth scale [4], giving the altitude of a moving celestial body



8 An orrery made in 1790 is illustrated here. The name originates from the Earl of Cork and Orrery, for whom an elaborate instrument was made Orreries indicate the movements of the planets round the Sun. In the orrery shown here the Sun is represented by a brass ball in the centre Round it move the three inner most planets. Mercury, Venus and the Earth; an ingenious system of gears makes the planets move round the Sun in the correct relative periods, even though on a scale of this kind it is naturally impossible to give the correct relative distances. The lands and seas of Earth are shown and also the Moon in its orbit round the Earth, which is inclined at the correct angle When by turning a handle, the mechanism is moved, the planets revolve round the Sun and the Moon revolves round the Earth More modern orreries are driven by clockwork, and in some the more distant planets than Earth and those discovered in the years after this model was made are also shown

Telescopes

The telescope is the main research instrument of astronomy. Without it our knowledge would be very limited indeed because other instruments – such as those based upon the principle of the spectroscope – depend upon telescopes to collect light that is to be analysed. George Ellery Hale (1868–1938), who was largely responsible for the building of the 200in (508cm) reflector at Palomar, in the United States, which was for many years the most powerful telescope in the world, once said that his call was always for "More light!" This is still true today, for modern astronomers are continually striving to investigate extremely faint objects that lie at immense distances from the Earth, and are gazing ever more deeply into the universe.

How refractors work

Telescopes are of two main kinds, refractors and reflectors [Key]. Each type has its own advantages and, unfortunately, its own drawbacks. Refractors, developed during the first decade of the seventeenth century, were first in the field and were used by pioneers such as Galileo (1564–1642). In a refractor the light from the object to be studied passes through a specially shaped lens known as an object-glass or objective; the rays of light are brought to focus and the resulting image is magnified by a second lens known as an eyepiece or ocular. The larger the object-glass the greater the light-grasp of the telescope; thus a 6in (15.2cm) refractor (that is to say, a refractor with an object-glass six inches across) is twice as powerful as a small 3in (7.6cm) refractor.

The only function of the object-glass is to collect light; all the actual magnification is done by the eyepiece. Every astronomical telescope is equipped with several eyepieces, which can be used as desired. The limit depends upon the amount of light available. Thus if, say, an eyepiece giving a magnification of 500 were used with a 3in (7.6cm) refractor, the resulting image would be so faint that it would be useless; to make use of a magnification of 500 a larger object-glass would be needed.

All refractors have one defect in common: they produce false colour. This is due to the nature of light itself, which is a blend of all the colours of the spectrum [1]. As the ray of light passes through the object-glass it is bent or refracted in order to be brought to focus, but the longer wavelengths are bent less sharply than the shorter ones. Thus the red part of the beam is bent less than the blue and so is brought to focus in a different place. The result is that a bright object such as a star is associated with false colour that may look beautiful but which, to the astronomer, is most unwelcome. This can be partly remedied by using compound object-glasses with one lens of crown glass and the other of flint glass; these have different refractive properties, and the false colour is reduced. It could be almost entirely removed (as it is in cameras) by adding more lenses, but that would significantly reduce the amount of light reaching the eye of the observer, and this is a fundamental consideration in astronomy [2].

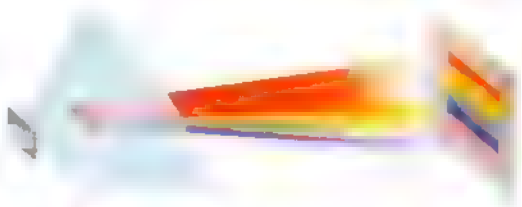
Reflecting telescopes

The reflecting telescope, of which the first working example was made by Isaac Newton (1642–1727) in about 1671, works on an

CONNECTIONS

See also

1A

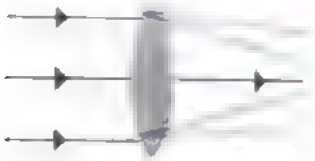


B

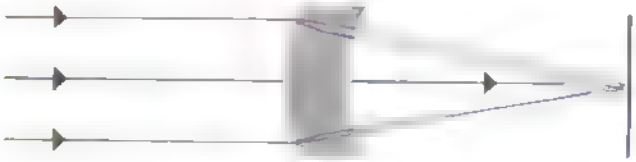
1 When white light, which contains all wavelengths of the visible spectrum, is passed through a glass prism, the beam is split up [A] (the colours bend unequally), into the spectrum ranging from the longest wavelength (red) to the shortest (violet). When one colour is passed through a hole in a screen and then through a second prism, there is no further splitting up [B]. By inverting the second prism the colours can be recombined [C].

2 The cause of the irritating false colour that is always present when a refractor is used is shown here. The light from the object passes through the object-glass and is split up to some extent so that the red rays are brought together from the lens at a different point from the blue rays [A]. The same is true if a different kind of lens is used [B]. The solution is to use a compound object glass [C] made up of two lenses combined, the errors then tend to cancel each other out and the false colour is appreciably reduced, although for colour correction a refractor is always inferior to a reflector.

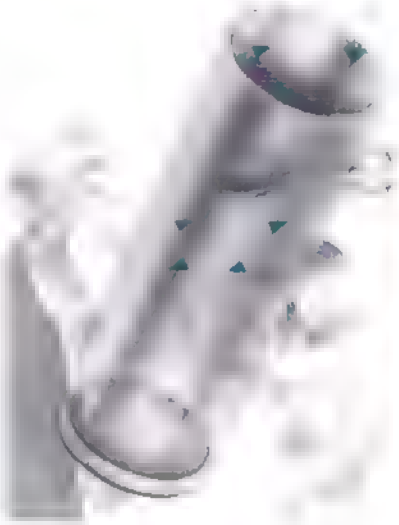
2A



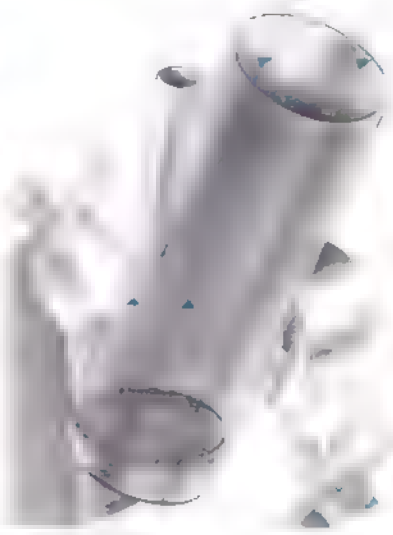
B



3



Newtonian reflector



Herschel's reflector

Cassegrain reflector

3 Reflectors are of various types. In the Newton an pattern the light is collected by a parabolic mirror and sent on to a flat mirror at an angle of 45 degrees, the light is sent into the side of the tube, where the image is formed and magnified. To avoid the admittedly small light loss due to the flat, William Herschel tilted the main mirror and dispensed with the flat. However, this design is unsatisfactory. In the Cassegrain the secondary is convex and the light is reflected back down through a hole in the main mirror. In some designs the returning light is diverted by a second flat mirror into the side of the tube, which avoids making a hole in the main mirror.

entirely different principle. On the Newtonian pattern [3] the light passes down an open tube until it hits a mirror at the far end. This mirror is curved; the shape is that of a paraboloid and the light is reflected back up the tube onto a second, flat mirror placed at an angle of 45 degrees. The light is then directed into the side of the tube, where it is brought to focus, and the image is magnified by an eyepiece as before. The presence of the flat mirror in the tube cuts out some of the light but the loss is not serious and with the Newtonian pattern there is no way of avoiding it.

Because a mirror reflects all colours equally there is no chromatic aberration – although a certain amount of false colour may be produced in the eyepiece. Modern mirrors are ceramic and are coated with a thin layer of some highly reflective substance such as aluminium or silver.

The Newtonian is not the only form of reflector. In the Cassegrain or Gregorian type [3] the second mirror is also curved and the light is reflected back through a hole in the main mirror. In the Herschelian type of reflector the main mirror is tilted and the secondary mirror is dispensed with altogether, but this involves distortions and as a result, Herschelian telescopes are now considered obsolete.

Advantages and disadvantages

Aperture for aperture a refractor is more effective than a reflector, but it is also more expensive because large lenses are harder to make than large mirrors. For this and other reasons all the world's largest telescopes are reflectors [5]. For the amateur astronomer the minimum really useful aperture is probably 3in (7.6cm) for a refractor and 6in (15.2cm) for a reflector.

The question of a mount is all-important. If the mounting is unsteady the telescope will be useless. It is highly desirable to use an equatorial stand in which the telescope is attached to an axis that is parallel to the axis of the Earth. With a driving mechanism added, the telescope can be driven in a way that compensates for the Earth's rotation and keeps the object under study perpetually in the field of view.

In a refractor [A] light from the object [1] passes through the lens to form an image [2]. The distance between the lens and the focal point is known as the focal length. Unless an extra lens system is used, the image is inverted.

In a reflector [B] light from the object [1] is collected by a curved mirror [3] and is brought to focus, forming an image [4].

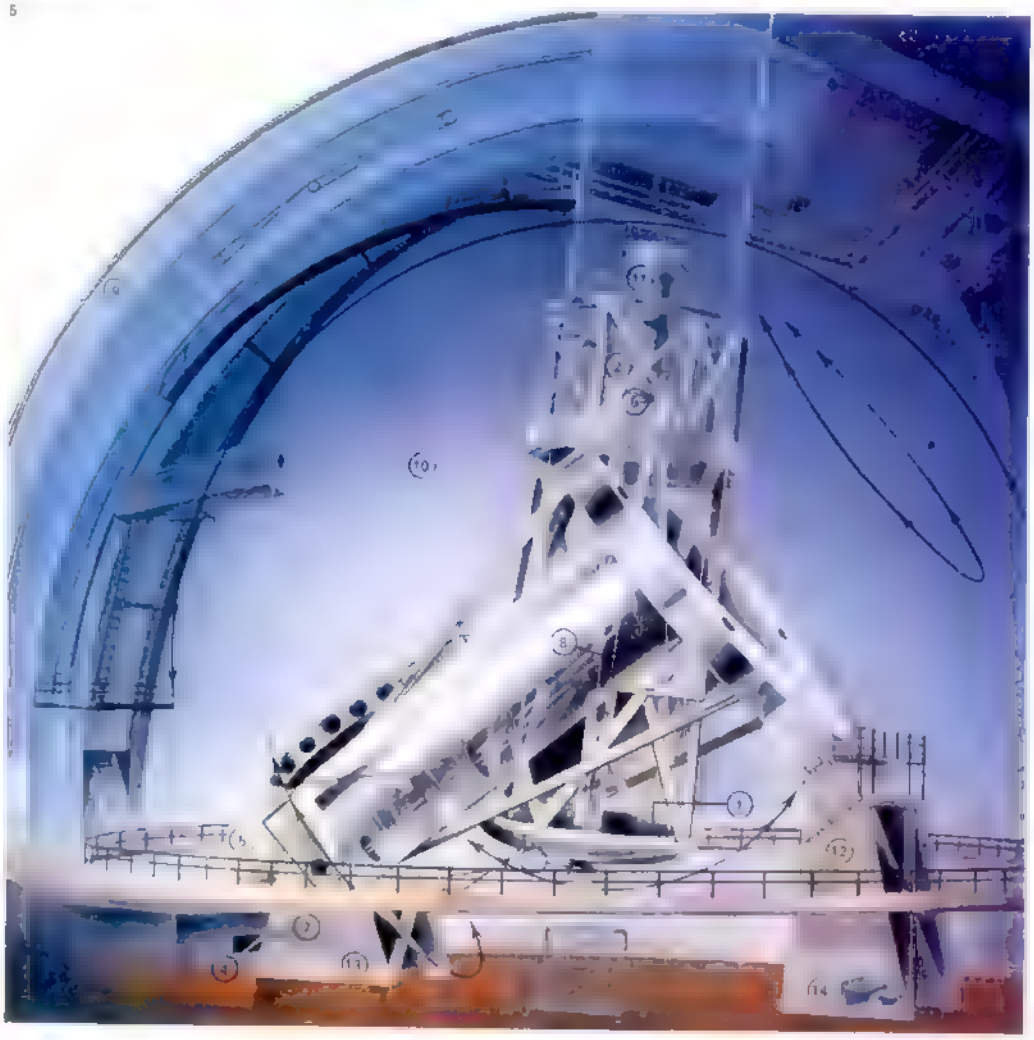
4 A more modern type of reflector is the Coudé, which has a secondary mirror and an extra, rotatable mirror on the polar axis of the telescope. Since the light rays are reflected in a constant direction the image formed is stationary and the observer need not move as the telescope rotates. This has the great advantage that heavy and delicate equipment can be set up and need not be moved. Most modern reflectors allow for a Coudé focus and can be used according to several optical systems, which permits great versatility; in some telescopes the change-over to a Coudé system can be made very quickly. It can also be used with refractors.

5 The 200in (508cm) Hale reflector was for many years the world's largest and was unrivalled in its light-grasp. The diagram shows the primary mirror [1], observer's cage [2], Cassegrain focus [3], the Coudé focus [4], the southern end of the polar axis [5], the Cassegrain and Coudé secondary mirrors [6], the right ascension drive [7], the declination axis [8], the dome shutter, with an opening of 9m (30ft) [9]; the 42m (137ft) dome [10], the primary focus, 16.5m (54ft) [11], the northern pillar [12], the southern pillar [13]; and the control panel [14], from which the telescope can be made to point to any part of the sky.

4

Coudé telescope

5



6 The cage for the observer can be set up with in the tube itself in a telescope that is as large as the Hale reflector. This means that photographs can be taken at the prime focus. This is an obvious advantage because it means that no secondary mirror is needed – and every reflection from a mirror inevitably involves loss of light. The amount of light blocked out by the observer's cage is tolerated because of the benefits the cage provides.

6

7 One of the observers at Palomar loads a plate at the 200in (508cm) reflector, which was here being used at the Coudé focus. To change from one optical system to another takes a certain amount of time, but the operation is routine.

7

Great observatories

It is often thought that an astronomical observatory consists simply of a dome shaped building containing a telescope. This is certainly true of some amateur observatories, but professional observatories are extremely elaborate and contain equipment of many kinds. Great care has to be taken in siting an observatory because a dark sky with no interference from artificial light and a clear, transparent atmosphere are absolutely paramount for good observation.

Today, little astronomical work is carried out visually – that is to say, by an observer sitting at the eyepiece of a telescope. Virtually all research is carried out by means of photography and the world's largest telescopes are used as giant cameras.

Observatory sites and equipment

To photograph a very faint object such as a remote galaxy it is necessary to make a time exposure, which may last many hours. Consequently, stray light is probably the astronomer's worst enemy. Today, with the spread of cities and the consequent marked increase in light pollution, it is becoming more and more difficult to find really good sites that combine darkness with a high percentage of clear, cloudless nights. In addition the Earth's atmosphere is subject to turbulence and also absorbs light, so that it is desirable to set major telescopes upon mountains, above the densest atmospheric layers. This means that an observatory must be virtually self-contained, with accommodation for the observers as well as workshops, photographic laboratories and lecture halls. Even so, the usual procedure is for the observer to spend a limited time at the observatory itself and then return to his office with the photographic results of his work.

Observatories built during the late nineteenth century were equipped with large refractors. Pride of place must go to the 40in (102cm) refractor at the Yerkes Observatory in the United States. It is not likely that any larger refractor will be built, because a lens has to be supported round its edge and above a certain limiting size (about 40in [102cm]) the lens starts to distort under its own weight. Large lenses are also subject to severe chromatic and spherical aberrations that distort the image; these aberrations can be avoided with mirrors. For these reasons most modern telescopes are reflectors.

The Hale Observatories

The most famous planner of giant telescopes was George Hale (1868–1938) who designed observatories and persuaded friendly millionaires to finance them. At Mount Wilson, in California, Hale was responsible for the erection of first a 60in (152cm) reflector and then one of 100in (254cm). The latter, completed in 1918, remained the largest telescope in the world for more than 30 years and was instrumental in making fundamental advances in astronomy. It was then surpassed in 1948 by the 200in (508cm) reflector at Palomar, California, also masterminded by Hale, who died before the telescope was completed. Fittingly, Mount Wilson and Palomar are now administered jointly under the name of the Hale Observatories. The 200in (508cm) telescope can be used on three optical systems (prime focus, Cassegrain and Coudé), and is so large that the observer's cage can be placed inside the tube

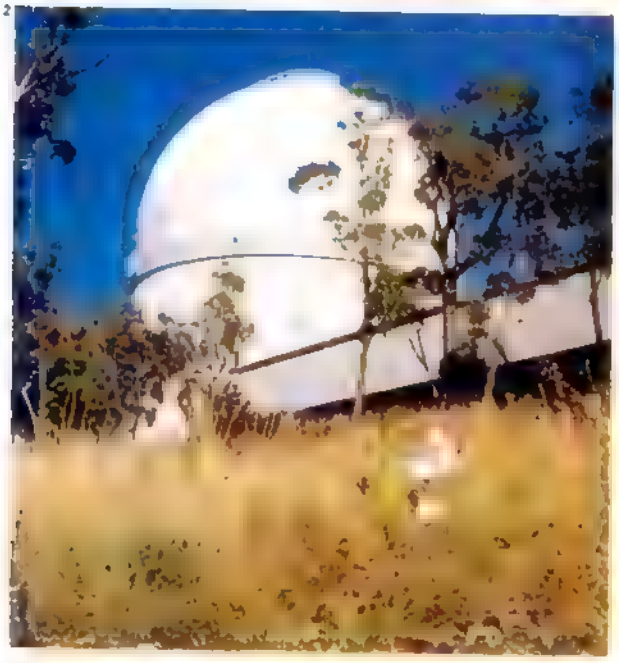
CONNECTIONS

See also

1

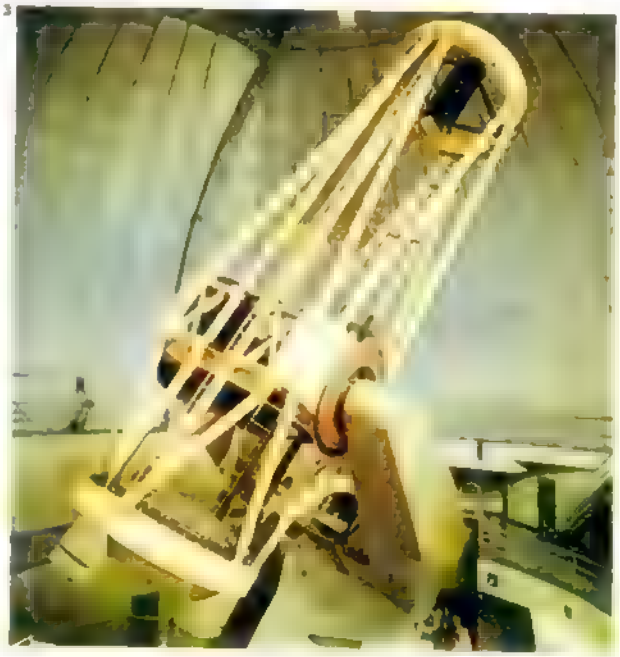
1 The world's largest telescope is the 236in (600cm) reflecting instrument at Zelenchukskaya, in the Northern Caucasus. Its technical advantage over the Hale 200in (508cm) at Palomar is considerable, although observing conditions in this region are not quite as good as those in California. The 236in instrument is entirely Soviet made. It has an alta-zimuth mounting with a more complicated drive mechanism than the usual equatorial mount. The dome is made to the conventional pattern. The telescope will be used mainly to study remote star systems, exploiting its immense light-gathering power. Initial tests were carried out in 1974.

2



2 The Anglo-Australian telescope at Siding Spring Mountain near Coonabarabran, New South Wales, is a 153in (389cm) reflector. Four different optical systems can be used: prime focus, f/8 or f/15 Cassegrain and f/36 Coudé. The total mass of the telescope is 326 tonnes. Its design is similar to the 150in (381cm) Kitt Peak telescope.

3



3 Lick Observatory in California was founded with a donation from James Lick in 1874–5, and came under the direction of the University of California in 1888. The principal instrument is the 120in (305cm) reflecting telescope, shown here, that went into operation in 1959. Many of the design features are similar to the 200in at Palomar.

– where the 45-degree flat mirror of an amateur Newtonian telescope is sited.

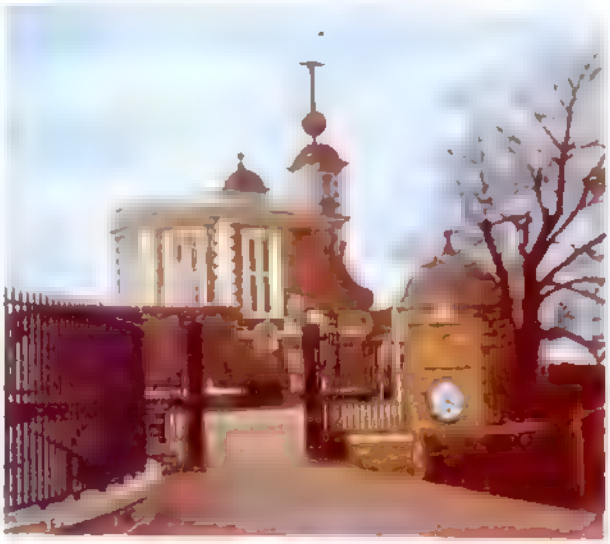
The 48in (122cm) Schmidt telescope [4] at Palomar is used purely for photography and incorporates a spherical mirror together with a complicated correcting plate. The advantage of a Schmidt telescope is that it can photograph wide areas of the sky with a single exposure, whereas the field of a telescope such as the 200in (508cm) is by its very nature extremely limited.

The world's largest telescopes

The 200in Palomar telescope is no longer the largest optical telescope in the world, since the Russians have built a 236in (600cm) reflector in Zelenchukskaya [1]. There are also various telescopes in the 100-160in (254-406cm) range. Major observatories have been set up in the Southern Hemisphere, where the skies are clear and the important objects of the far south, such as the Magellanic Clouds, are accessible. There are sites in Australia [2], South America and South Africa, where the main telescopes of the Republic have been collected together at a single site – Sutherland, in Cape Province – for particularly good observing conditions.

The largest telescope ever set up in Britain is a 98in (249cm) reflector, at Herstmonceux, Sussex, known as the INT or Isaac Newton Telescope. During the 1950s the instruments at the famous observatory at Greenwich [Key] were moved to Sussex but even there difficulties arose with cloud and scattered light and the 98in (249cm) is to be moved to a new site in the Canary Islands, where it will be joined by a new 160in (406cm) reflector.

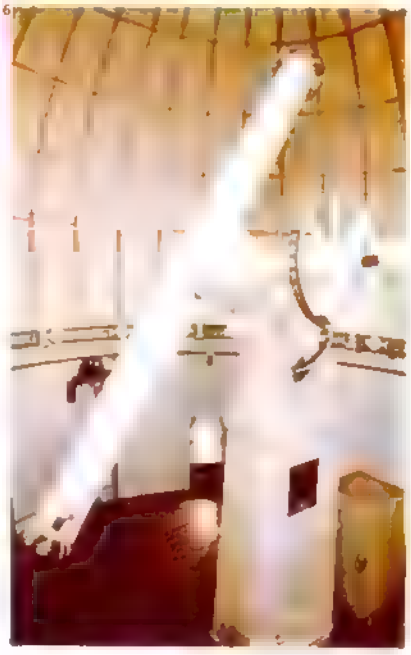
Some observatories have special roles, for instance at Kitt Peak in Arizona there is elaborate equipment for studying the Sun, while the Lowell Observatory in Arizona [5] specializes in planetary work. New observatories are now being planned, many conceived primarily to overcome the limiting effects of the Earth's atmosphere. Thus several artificial satellites have been equipped with telescopes, culminating in the achievements of Skylab – the first manned orbiting observatory – and its Soviet counterpart, Soyuz.



Flamsteed House, in Greenwich Park, London, designed by Christopher Wren (1632–1723), is the site of the old Royal Observatory, set up in 1675. The instruments have been moved to Sussex and the old observatory is now a museum.

4 Modern telescopes have immense light gathering power. But the conventional telescope is able to cover only a small area of sky with one photographic exposure. For studying individual objects such as galaxies this does not matter, but it means that to compile a photographic map of the whole sky would take too long. The principle of the Schmidt telescope [B], developed by Estonian optician Bernhard Schmidt (1879–1935) in 1932, enables large areas to be photographed with each exposure. There is a spherical mirror and a glass correcting plate over the end of the tube to compensate for optical distortion. The light passes through the plate [1], to the mirror [2] and is reflected onto a curved photographic plate [3] in the tube. The Schmidt telescope at Palomar [A] has a 48in correcting plate.

5 The Lowell Observatory at Flagstaff, Arizona, was set up by Percival Lowell (1855–1916) in 1895 mainly to study Mars; the observatory has been known for its planetary work, although much equipment, including a large reflector has been added since Lowell's time. The photograph shows the dome of the 24in (61cm) refractor that Lowell used for his Martian studies from 1895 to 1916. The fine-quality optics are as good as when they were new.



6 The 26in (66cm) telescope at Washington, DC, was one of the earliest of the great refractors. It was installed in 1862 and with it Alvan Clark (1832–97), who ground the object-glass, found the white dwarf companion of Sirius. Asaph Hall (1829–1907) used it to discover Phobos and Deimos, the two satellites of Mars in 1877. The photograph shows it as it is today with the telescope balanced by a counterweight, using a German type mounting.

Invisible astronomy

Up to the 1920s astronomers had to depend entirely upon the visible light coming from objects in space. This was a severe limitation because visible light makes up only a small part of the whole range of wavelengths or "electromagnetic spectrum".

Light may be regarded as a wave motion and the colour of the light depends on its wavelength. The usual unit of wavelength is the ångstrom, which is equal to one hundred millionth of a centimetre. Visible light extends from about 4,000 Å for violet light up to 7,200 Å for red. If the wavelength lies outside these limits the "light" does not affect our eyes. Below the violet end of the visible spectrum come ultra-violet, X rays and the very short, penetrating gamma rays; beyond the red end there are infra-red, microwaves and finally radio waves, whose wavelengths may amount to many kilometres.

Radio waves from space

The discovery of radio waves from space was made by Karl Jansky (1905-50) in the United States in 1931. The discovery was fortuitous. Jansky, a radio engineer, was investigating the nature of static when he found that he was picking up emissions from the sky. He tracked them down to the Milky Way. He published a few papers, but never followed the subject through. Before World War II an American, G. Reber, set up a dish-shaped radio telescope and made the first radio map of the Milky Way. During the war a British team led by J. S. Hey found that radar equipment was being jammed not by transmissions from Germany, as was first thought, but by radio waves from the Sun.

Subsequently, radio telescopes were set up and a new branch of science was well under way. It was found that the Sun is a radio source, but by cosmic standards not a powerful one; it is obtrusive only because it is so close to the Earth. Jupiter is also known to be a source of radio waves. But most radio sources lie far beyond the Solar System. Those in our Galaxy include many supernova remnants, of which the Crab Nebula is the most celebrated example. At greater distances still are the radio galaxies, which are extremely powerful at long wavelengths, although why is still not known.

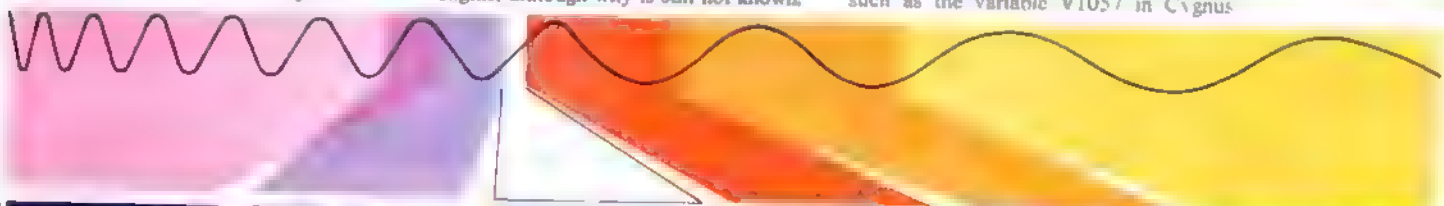
Radio astronomy has added tremendously to our knowledge of the universe. Without it little would be known about pulsars, which are neutron stars, or quasars, which are extragalactic and may well be the most powerful objects known to man. Moreover, radio waves have been studied from greater distances than visible light waves, so that our information about the most distant regions of the universe is derived entirely from radio work.

Infra-red radiations

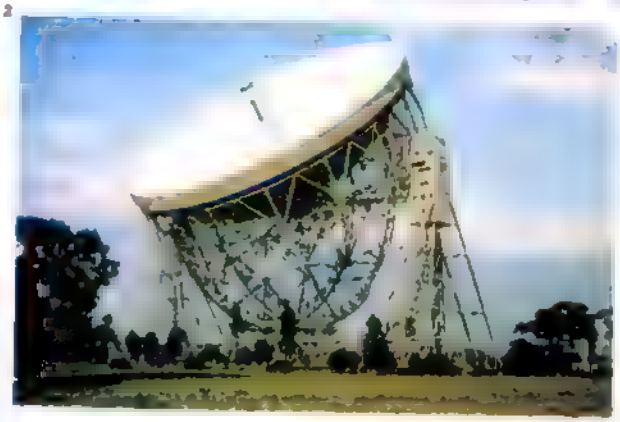
Beyond the longwave end of the visible range is the infra-red region of the electromagnetic spectrum. Most infra-red radiations are absorbed by the upper atmosphere and these are studied by means of equipment carried by satellites. However, there are a few "windows" through which infra-red radiations penetrate the atmosphere and these can be studied from the ground. This branch of research has provided a great deal of information about stellar evolution. For instance, there are extremely young stars such as the variable V1057 in Cygnus

CONNECTIONS

See also

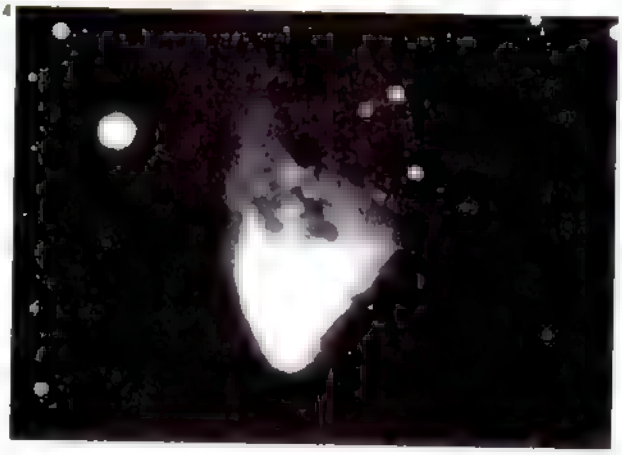


1 The electromagnetic spectrum shows the restricted "windows" in which radiations from space can reach Earth. Many of the largest wavelengths are blocked out, as are all the shortest. The illustration is not to scale.



2 The 76m (250ft) paraboloid at Jodrell Bank was for many years the world's largest fully steerable radio telescope. It was planned by the director, Sir Bernard Lovell (1913-), and has been responsible for fundamental advances in our knowledge. In its earliest days it was used for tracking artificial satellites and space probes, but it is now used exclusively for research into the stars and the galaxies.

3 The Arecibo radio telescope, in Puerto Rico, has been built in a natural bowl. It is 300m (1,000ft) in diameter. Although the radio telescope is not steerable, some degree of direction can be obtained by moving the receiving aerial.



4 Hubble's variable nebula in Monoceros is 6,500 light years away and is associated with a variable star, R Monocerotis. The infra-red radiation of which is its source.

which seem to be surrounded by dust clouds, so that the dust heated by the star contained in the cloud sends out excess infra-red. There are even some objects that are detectable only by infra-red techniques, for example Brecklin's Object inside the Orion Nebula.

Ultra-violet, gamma- and X-rays

Ultra-violet radiation, X-rays and gamma rays lie beyond the shortwave end of the visible band of the electromagnetic spectrum. For high-energy ultra-violet, X ray and gamma-ray studies (that is, radiations below 2,900A), equipment carried by rockets or satellites has to be used since these radiations are absorbed by the upper atmosphere.

Ultra-violet satellites, such as Copernicus and the International Ultra-violet Explorer, have studied radiation from the rare very hot stars – at temperatures of 30,000 C (54,000 F) or more. This ultra-violet light is affected by the gas clouds it passes through, and detailed study has revealed the proportions of different chemical elements in the gas in space.

Many powerful X-ray sources have been detected by X-ray astronomy satellites, for example Uhuru (launched in 1970), HEAD-1 and the Einstein Observatory. Most are double stars in our own Galaxy, where one member is a very compact neutron star which is pulling gases off its ordinary companion: the hot gas stream emits X rays copiously.

The sophisticated Einstein Observatory (1978-81) took the first-ever pictures of X ray sources, revealing the distribution of extremely hot X-ray emitting gas in the remains of exploded stars, and between the galaxies in very distant clusters of galaxies.

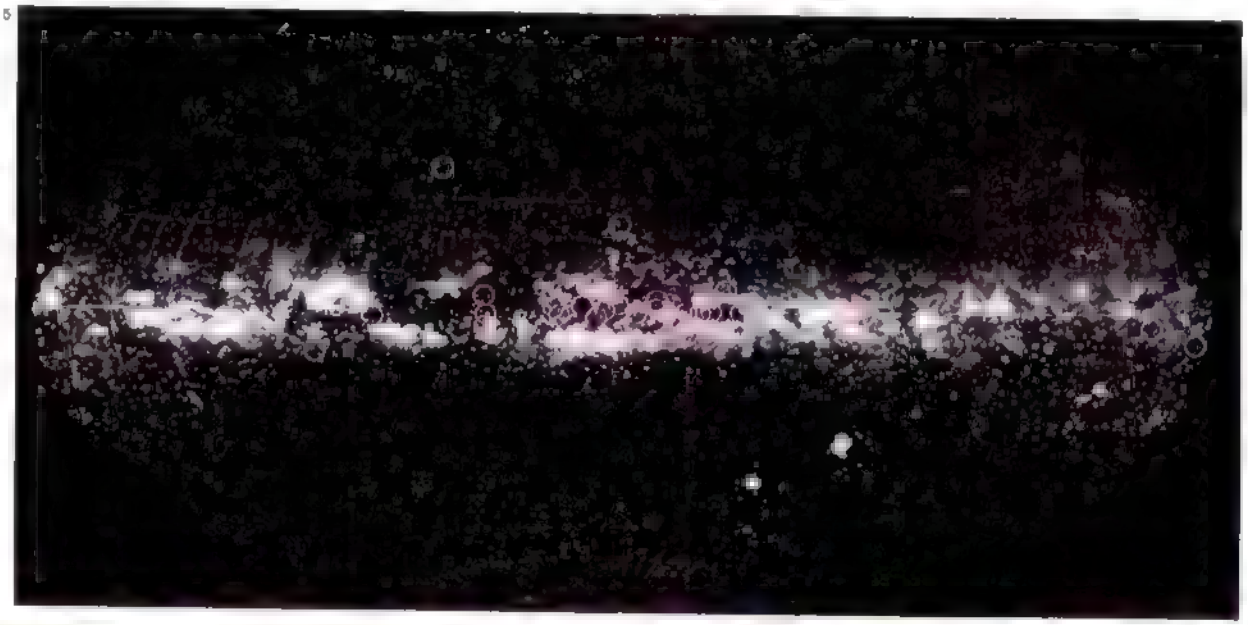
Gamma-ray observatories have found radiation coming from giant gas clouds in our Galaxy. In addition, gamma-ray satellites launched for military surveillance of the Earth have discovered very powerful, short bursts of gamma-rays from space.

Modern knowledge of the universe relies as much on "invisible astronomy" as on traditional optical astronomy. Indeed, astronomers now regard light from space as no more important than other radiation.

KEY

An improvised aerial was set up by Karl Jansky in 1931. He intended to study static on behalf of Bell Telephone and in the process discovered radio waves from the Milky Way, which led to the modern science of radio astronomy.

5

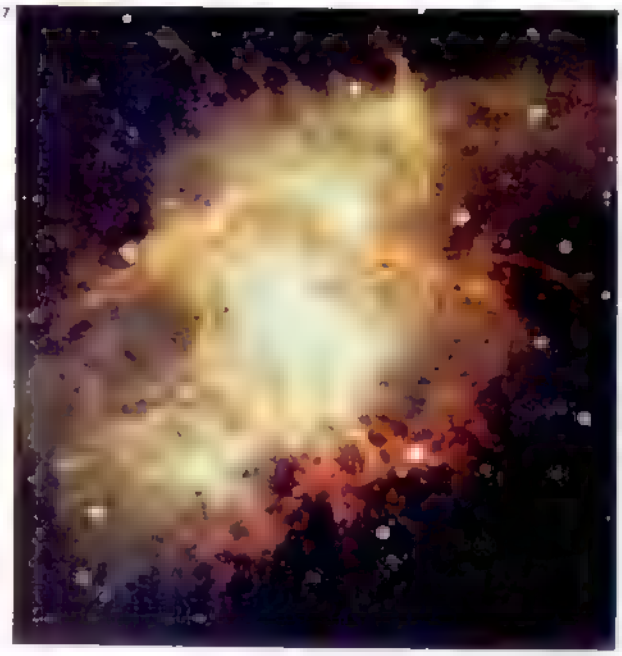


5 About 160 locations of X ray sources along the main plane of the Milky Way have now been identified and the distribution indicates that these sources belong to our own Galaxy. More sources are being found each year. Not all are permanent: the British Ariel satellite found highly energetic sources in Taurus near but not associated with the Crab Nebula, which lasted for some months in 1974. The Uhuru satellite discovered X-ray pulsars, which give regular pulses of radiation every few seconds; the Dutch Ariel satellite found bursters, sources with rare powerful outbursts.

6

6 The Crab Nebula is an intense source of X-rays as well as all other radiations; an X-ray view from the orbiting Einstein Observatory shows X-radiation from the electrons in the Crab's magnetic field. The Crab Nebula is the remnant of a supernova seen by Eastern Astronomers in AD 1054, and the small intense X-ray source is the Crab Pulsar, the rapidly spinning neutron star left at the centre of the enormous explosion.

7



7 An ordinary photograph of the Crab shows its outer crab-like filaments. Although it lies an immeasurable 6,000 light years away, the Crab can be seen with a small telescope. It lies in the constellation of Taurus.

Evolution of the Solar System

The question of how the Earth came into existence is one that has intrigued mankind for centuries. It was not until comparatively recently that plausible theories were advanced and even today it is impossible to be sure that the main problems have been solved, but at least some concrete facts exist

Fallacies disproved

The concept of a central Earth, with the Sun moving round it, was abandoned during what is often called the "Copernican revolution", which began in 1543 with the publication of Copernicus's book *De Revolutionibus Orbium Caelestium* and was finally completed by the work of Newton in the latter part of the following century. Therefore it could be assumed that the Earth must have been formed in the same way as the other planets rather than being a special case

Several centuries ago Archbishop Ussher of Armagh maintained, on religious grounds that the world had come into existence at a definite moment in the year 4004 BC Geological evidence soon disproved this but it was not until much more recently that any reliable estimates could be made. The modern estimate for the age of the Earth is between 4,500 and 4,700 million years, or 4.5–4.7 aeons (one aeon being equal to a thousand million years), and this figure is as reliable as modern knowledge permits

Further confirmation of the Earth's age has been obtained from analyses of the rocks brought back from the Moon by the Apollo missions and the Soviet automatic probes, it is now known that the Moon and the Earth are about the same age and no doubt the same is also true about the other planets. The Sun must be at least as old as the planets and probably rather older There can be no doubt that the Sun has been responsible for the origin and formation of the whole of the Solar System

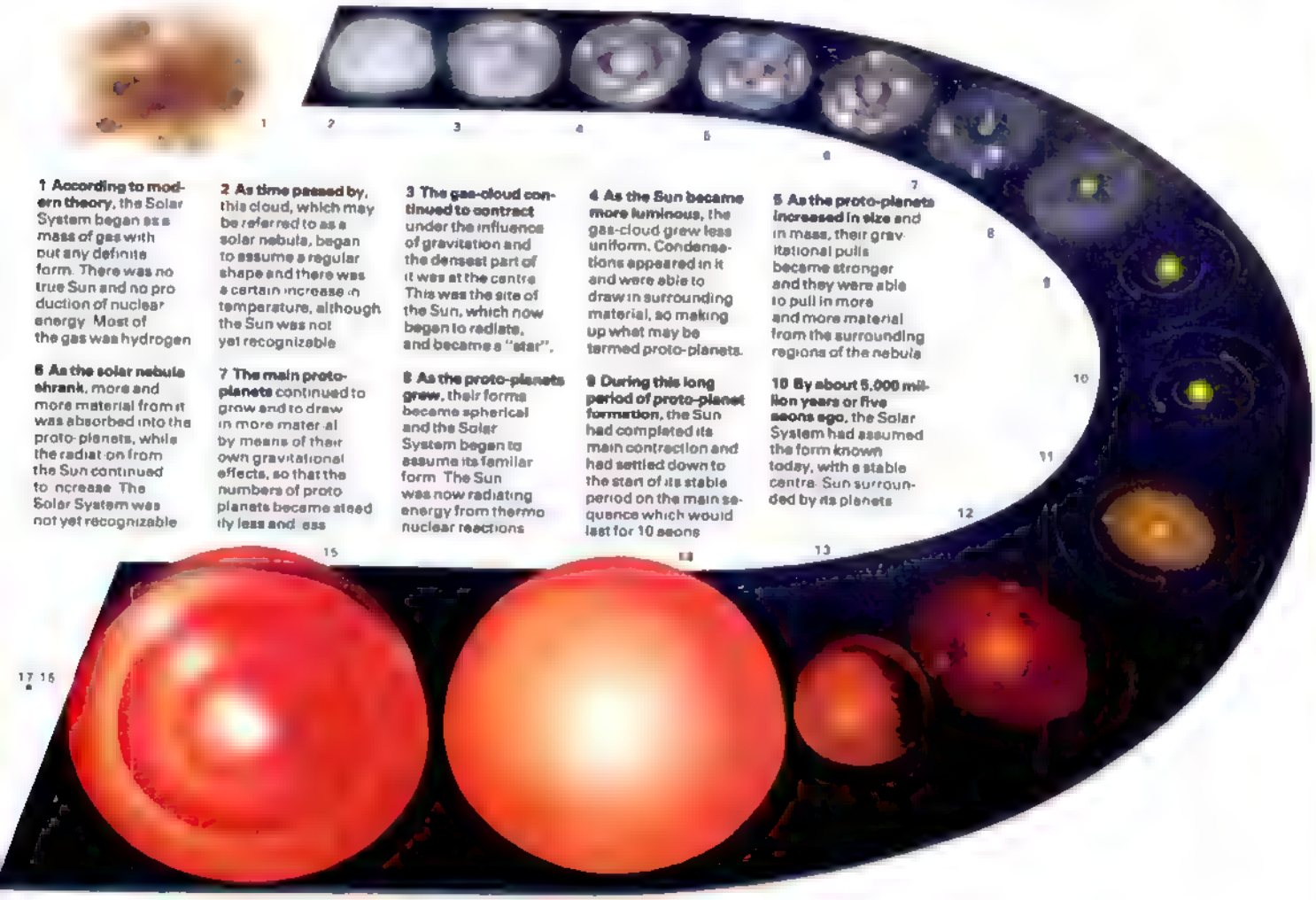
The first scientific theories

The first serious attempt to explain the origin of the Solar System scientifically was made by the French mathematician Pierre Laplace (1749–1827) in 1796 (although earlier ideas, less purely scientific, had been proposed by Thomas Wright [1711–86] in England and Immanuel Kant [1724–1804] in Germany) According to Laplace, whose "nebula hypothesis" [Key] elaborated an idea proposed by René Descartes (1596–1650) in 1644, the planets were formed from a rotating gas-cloud that shrank under the influence of gravitation As it contracted, the cloud shed various rings, each of which condensed into a planet. The theory would mean that the outermost planets were the oldest and the innermost planets the youngest, with the Sun itself representing the remaining part of the original gas-cloud

The nebular hypothesis was accepted for many years, but it was then found to have basic mathematical weaknesses and was abandoned. Next came a number of tidal theories, including the ideas proposed in America by Thomas Chamberlin (1843–1928) and Forest Moulton (1872–1952), who revived George de Buffon's original idea (1745), developed by James Jeans (1877–1946) in England [19] It was assumed that the planets were formed by the action of a passing star, which came close to the Sun and pulled off a vast tongue of

CONNECTIONS

See also



1 According to modern theory, the Solar System began as a mass of gas without any definite form. There was no true Sun and no production of nuclear energy Most of the gas was hydrogen

2 As time passed by, this cloud, which may be referred to as a solar nebula, began to assume a regular shape and there was a certain increase in temperature, although the Sun was not yet recognizable

3 The gas-cloud continued to contract under the influence of gravitation and the densest part of it was at the centre This was the site of the Sun, which now began to radiate, and became a "star".

4 As the Sun became more luminous, the gas-cloud grew less uniform. Condensations appeared in it and were able to draw in surrounding material, so making up what may be termed proto-planets.

5 As the proto-planets increased in size and in mass, their gravitational pulls became stronger and they were able to pull in more and more material from the surrounding regions of the nebula

6 As the solar nebula shrank, more and more material from it was absorbed into the proto-planets, while the radiation from the Sun continued to increase The Solar System was not yet recognizable

7 The main proto-planets continued to grow and to draw in more material by means of their own gravitational effects, so that the numbers of proto planets became steadily less and less

8 As the proto-planets grew, their forms became spherical and the Solar System began to assume its familiar form The Sun was now radiating energy from thermonuclear reactions

9 During this long period of proto-planet formation, the Sun had completed its main contraction and had settled down to the start of its stable period on the main sequence which would last for 10 aeons

10 By about 5,000 million years or five aeons ago, the Solar System had assumed the form known today, with a stable centre Sun surrounded by its planets

11 In perhaps 5,000 million years from now the Sun will have exhausted its supply of available hydrogen and its structure will change The core will shrink and the surface will expand considerably, with a lower surface temperature

12 The next stage of solar evolution will be expansion to the red giant stage, with luminosity increased by 100 times. The size of the globe will increase, with the overall increase in energy output, and the inner planets will certainly be destroyed

13 With a further rise in core temperature, the Sun will begin to burn its helium causing a rapid rise in temperature and increase in size The Earth can hardly hope to survive this phase of evolution as the Sun expands to 50 times its size

14 By now the Sun will be at its most unstable, with an intensely hot core and a rarefied atmosphere The helium will begin to burn giving the so called helium-flash. After a temporary contraction the Sun will be 400 times its present size

15 Different kinds of reactions inside the Sun will lead to an even greater increase of core temperature. The system of planets will no longer exist in the form we know today, but the supply of nuclear energy will be almost exhausted

16 When all the nuclear energy is used up, the Sun will collapse, very rapidly on the cosmic scale, into a small, dense and very feeble white dwarf It will continue to shine because it will still be contracting gravitationally

17 The final stage of the Sun will be that of a black dwarf, devoid of any light or heat, still circled by its surviving but now dead planets Black dwarfs may be common in the universe, but because they emit no radiation cannot be detected

material. As the star receded, the tongue of matter was left whirling round the Sun, and broke up into drops, each drop becoming a planet. This was in agreement with the sizes of the planets, since the giants (Jupiter and Saturn) lie in the middle part of the system, where the thickest part of the cigar-shaped tongue would have been.

However, this theory too has serious mathematical objections and few modern astronomers support any form of tidal theory for the evolution of the Solar System.

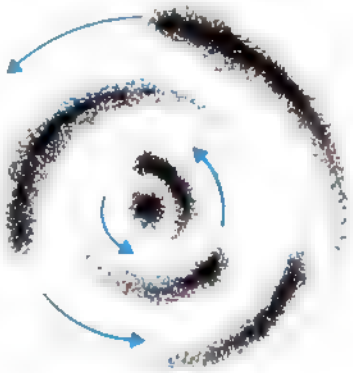
It was also proposed, by Fred Hoyle (1915–) that the Sun used to be a binary star [18] and that the companion exploded as a supernova, producing scattered debris which formed the planets before the companion itself departed by a kind of recoil action. This has, however, met with little support among astronomers.

The future of the Solar System

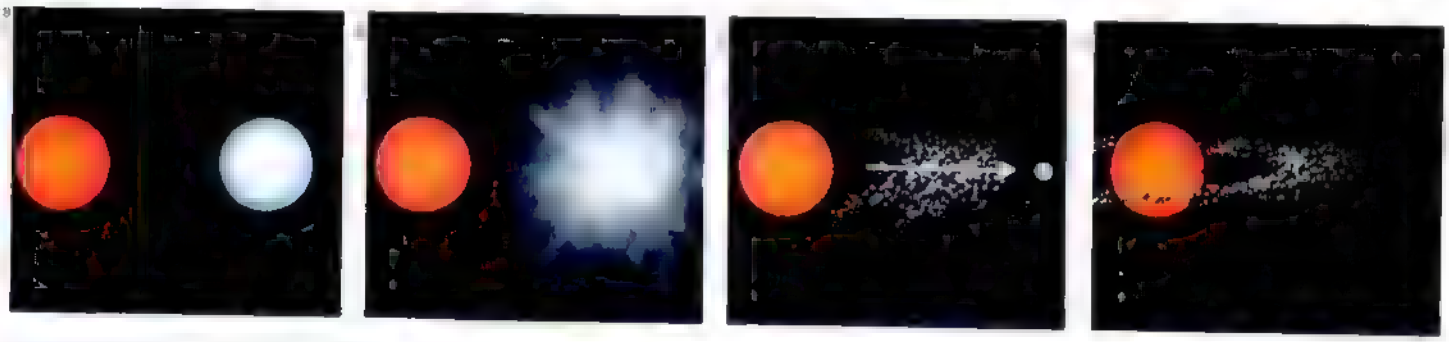
Modern theories assume the existence of what may be termed a solar nebula, which contained the material that gradually built up into the planets by an accretion process. The exact details are still a matter for debate, but in essence the theory seems to be valid; if so, the Sun and the planets have a common origin and are made of the same material.

At the moment the Sun is a stable main sequence star, but it will not remain so indefinitely. In the far future – perhaps in 5,000 million years or so – it will have to change its structure, since the supply of available hydrogen "fuel" will be exhausted. What will happen is that the Sun will expand into a red giant star, and there will be a period when it will send out about 100 times as much energy as it does at present.

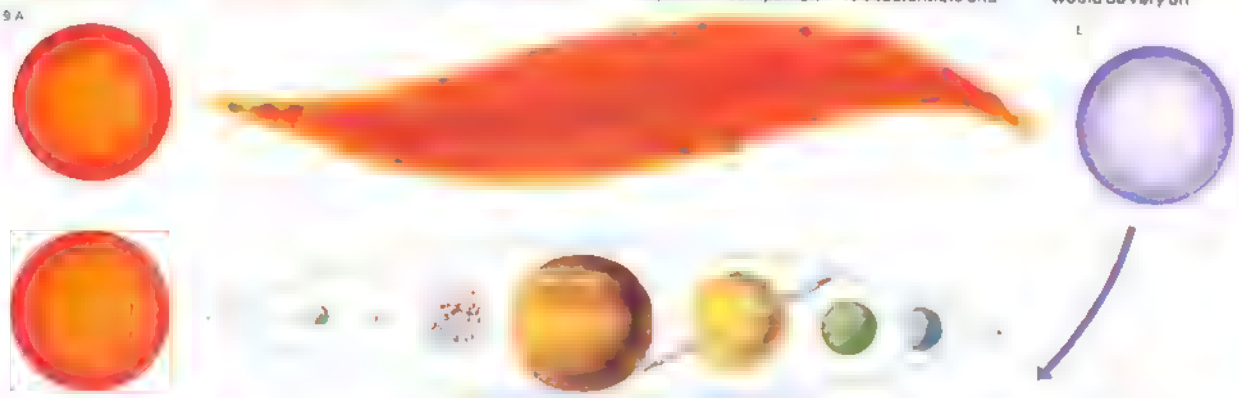
The effects of this expansion on the inner planets will be disastrous; even if they are not destroyed they will be stripped of their atmospheres and will become intolerably hot. Subsequently the Sun will collapse into a very small, feeble white dwarf star, still surrounded by the surviving members of its planetary system. The exact time scale is still a matter for debate, but one thing is certain: life on Earth cannot continue indefinitely and the Solar System in its present form must have a limited existence.



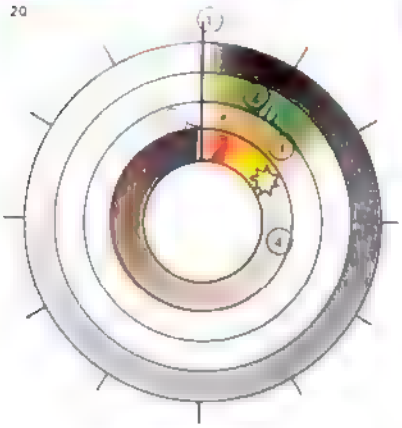
The nebular hypothesis, proposed by Laplace, assumed that before the birth of the planets the Solar System consisted of a gas-cloud which shrank, because of gravitational forces. This resulted in an increase in the speed of rotation and a ring separated from the nebula, the ring slowly condensed into a planet. Further rings were then thrown off, each producing a planet. The theory is mathematically weak.

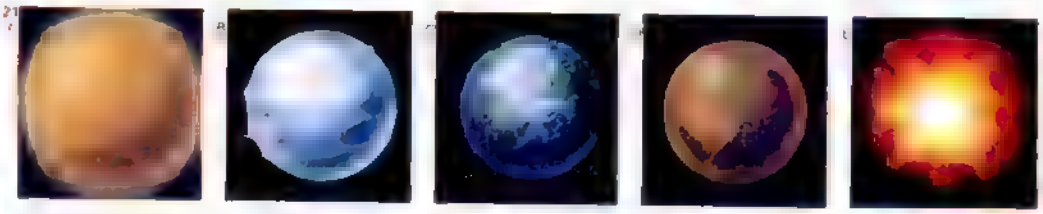


18 The binary theory of the origin of the Solar System was proposed by Fred Hoyle. He argued that the Sun once had a binary companion [A] which exploded as a supernova [B] and was blown off leaving a cloud of fragments [C] in orbit round the Sun, these fragments collected together by the process of accretion to form the planets [D], while the remnant of the supernova companion moved away into space and cannot now be identified. This theory is very difficult to substantiate and is not now generally favoured. If valid, it would mean that planetary systems would be very uncommon although on the modern solar cloud theories they are likely to be extremely common in our Galaxy.



19 The tidal theory, as proposed by James Jeans, is shown here. The Sun [A] is approached by another star [B] which pulls a tongue of matter off the solar surface. After the wandering star recedes, the tongue of matter breaks up into drops which form planets revolving round the Sun. Jupiter, the largest planet, is in the position of the thickest part of the cigar-shaped tongue.





20 The time scale of the Solar System can be represented on a 12-hour clock. From the inner circle outwards the life-span of the Sun, inner planets, Earth and outer planets are traced respectively. At the 12-o'clock position [1] the Solar System was created. After 4,000 million years conditions on Earth are favourable for life [2]. As a red giant the Sun engulfs the inner planets [3] to collapse as a white dwarf [4] and finally end its life as a black dwarf [5].

21 The life-span of the Earth started from the material of the solar nebula [A] which at first had no regular form. When it reached its present size [B] the original hydrogen atmosphere had already been lost and had been replaced by a new one, caused by gases sent out from the interior. Life could begin and today the Earth is moving in a stable orbit round a stable star, so that it is habitable [C]. This state of affairs will not persist indefinitely. When the Sun enters the red giant stage, the Earth will be over heated, the oceans will boil and the atmosphere will be driven off [D], finally the Earth will be destroyed [E].

Members of the Solar System

The Solar System is made up of one star – the Sun – nine principal planets and various bodies of lesser importance, such as the satellites that attend some of the planets. It is entirely dependent on the Sun, which is by far the most massive body and the only one to be self-luminous. The remaining members of the Solar System shine by reflected sunlight and appear so brilliant in our skies that it is not always easy to remember that in the universe as a whole they are not nearly as important as they look.

Groups of planets

The planets are divided into two well-marked groups. First come four relatively small planets: Mercury, Venus, the Earth and Mars, with diameters ranging from 12,756km (7,926 miles) for the Earth down to only 4,880km (3,032 miles) for Mercury. These planets have various factors in common. All, for example, have solid surfaces and are presumably made up of similar materials, although the Earth and Mercury are more dense than Mars and Venus.

Their orbits do not in general depart much from the circular, although the paths of Mercury and Mars are considerably more eccentric than those of the Earth and Venus. Mercury and Venus are known as the "inferior planets" because their orbits lie inside that of the Earth; they show lunar-type phases from new to full and remain in the same region of the sky as the Sun. Mercury and Venus are unattended by any satellites. The Earth has one satellite (our familiar Moon) while Mars has two, Phobos and Deimos, both of which are very small and different in nature from the Moon.

Beyond Mars comes a wide gap, in which move thousands of small worlds known as the asteroids, planetoids or minor planets. Even Ceres, the largest, is only about 1,000–1,200km (600–750 miles) in diameter. This is much larger than was once thought, but is still small by planetary standards. It is not therefore surprising that the asteroids remained undiscovered until relatively recent times. Ceres was discovered in 1801, only one of this multitude of asteroids, Vesta, is ever visible from Earth without the aid of a telescope.

Far beyond the main asteroid zone come the four giant planets Jupiter, Saturn, Uranus and Neptune. These worlds are quite different from the terrestrial planets: they are fluid (that is, gas or liquid) rather than solid bodies with very dense atmospheres. Their masses are so great that they have been able to retain much of their original hydrogen; the escape velocity of Jupiter, for instance, is 60km (37 miles) per second as against only 11.2km (7 miles) per second for Earth. Their mean distances from the Sun range from 778 million km (483 million miles) for Jupiter out to 4,497 million km (2,794 million miles) for Neptune. Conventional diagrams of the Solar System tend to be misleading as far as scale is concerned, it is tempting, for example, to assume that Saturn and Uranus are lying next to each other when in fact the distance of Uranus from the Earth's orbit is about twice that of Saturn.

The giant planets compared

The giant planets have various points in common, but differ markedly in detail. Their densities are comparatively low and the

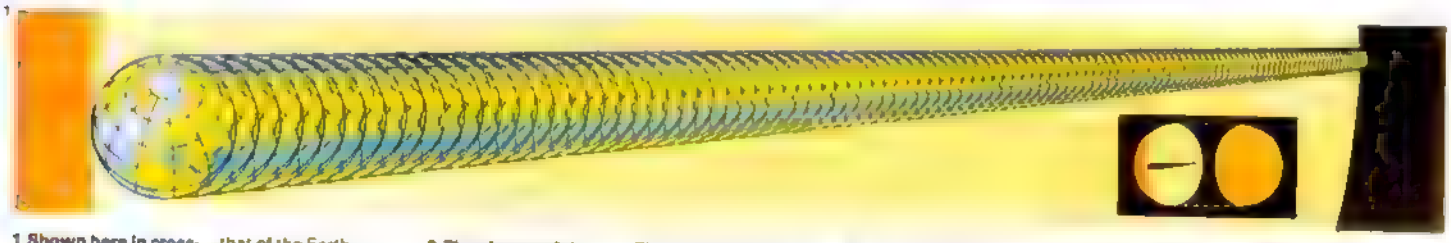
CONNECTIONS

See also

The Moon

The planet Mercury

Mars: planet



1 Shown here in cross-section, the Sun has an equatorial diameter 109 times that of Earth, or 1,392,000 km (865,000 miles). Despite the fact that its volume is more than a million times that of the Earth, its mass is only 333,000 times that of the Earth, because the density is lower. (The mean specific gravity, on a scale where water = 1, is only 1.4.)

2 The planets of the Solar System are shown to the same scale. On the right is a segment of the Sun [1]; from its surface rises a huge prominence [2] made up of glowing gas. Then come the inner planets: Mercury [3], Venus [4]; the Earth [5] with its Moon [6], and Mars [7]. Mars has two dwarf satellites Phobos [8] and Deimos [9] (exaggerated here). If shown to the correct scale, they would be too small to be seen without a microscope. Then come the asteroids [10], of which even the largest is only about 1,000–1,200km (600–750 miles) in diameter. Beyond lie the giant planets Jupiter [11] with its four large satellites Io [12], Europa [13], Ganymede [14] and Callisto [15]; two smaller moons are also seen in the picture. Saturn [16] with its retinue of satellites, of which the largest is Titan [17], Uranus [18] with five satellites, Neptune [19] with its large satellite Triton [20], and finally Pluto [21].

density of Saturn is actually less than that of water. Although Jupiter is seen soleby by reflected sunlight, the planet does generate some heat of its own. However, even though the core temperature must be high, it is not nearly high enough for nuclear reactions to begin, so that Jupiter cannot be compared to a star like the Sun

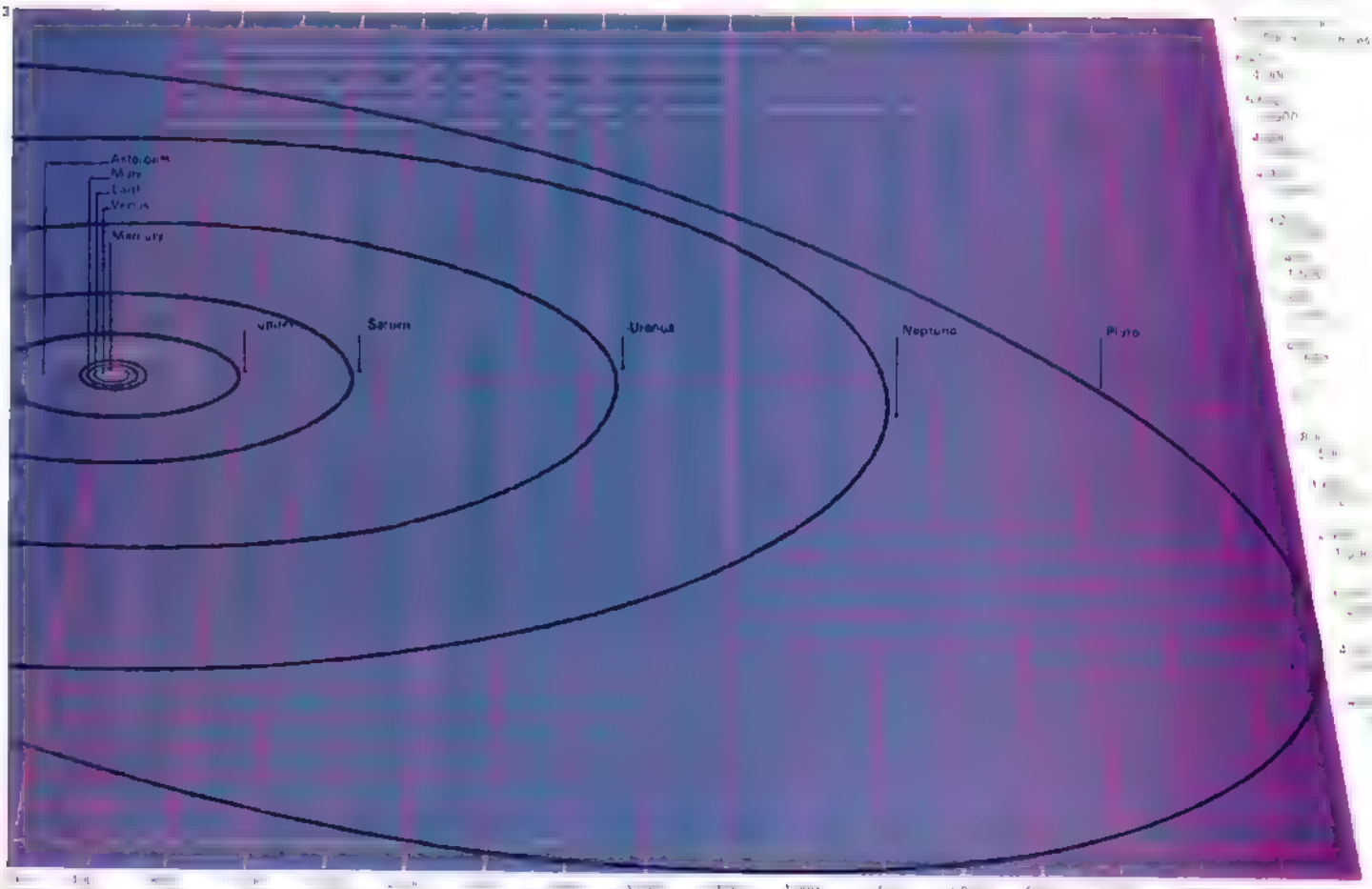
The outer planets

Five of the planets – Mercury, Venus, Mars, Jupiter and Saturn – have been known from ancient times, since all are prominent naked eye objects. Uranus, which is just visible with the naked eye, was discovered fortuitously in 1781 by William Herschel (1738–1822), and Neptune was added to the list of known planets in 1846, as a result of mathematical investigations carried out concerning the movements of Uranus. All the giants are attended by satellites, Jupiter has 16 attendants, Saturn 17, Uranus five and Neptune two. Several of these attendants are of planetary size with diameters at least equal to that of Mercury

The outermost known planet is Pluto, discovered in 1930 by astronomers at the Lowell Observatory, Flagstaff, Arizona. It is not a giant, being smaller than the Moon, and is attended by a satellite, Charon, one-third the diameter of Pluto itself. Whereas most of the planets have orbital inclinations similar to that of the Earth (the difference is 7 degrees for Mercury, much less for the remainder), the orbit of Pluto is tilted at the relatively steep angle of 17 degrees and the orbit is so eccentric that at perihelion, or closest approach to the Sun, Pluto will come closer in than Neptune. Pluto seems, in fact, to be in a class of its own, and it may be a former satellite of Neptune which has achieved independence. Whether some more distant planets exist beyond the orbit of Pluto is a challenge to technology

Comets are also members of the Solar System. They contain both dust particles and volatile material together with tenuous gas, most of them have eccentric orbits. Finally there are a great number of meteroids, which may be regarded as the débris of the Solar System, some meteors are certainly associated with comets

The Sun is an ordinary main sequence star with a magnitude of +5. It is the body on which the Solar System depends and its volume is more than a million times greater than that of the Earth. It is, in fact, far more massive than all the planets combined. However, the Sun is small when compared with a giant star. The diagram shows the Sun alongside a segment of the red giant star Betelgeuse which marks Orion's right shoulder. Betelgeuse is of spectral class M2 – a very cool star – but has an absolute magnitude of −5.5. Its diameter is 300 400 times that of the Sun and its globe is large enough to contain the Earth's orbit



3 The map of the Solar System shows the approximate orbital inclination of each of the nine planets against a grid giving distances in kilometres and miles. The following measurements for diameter and rotation period refer to each planet's equator. The sidereal period is the time taken by each planet to orbit the Sun once

Mercury
Rotation
58
Mass 0.05
Surface gra
Escape velocity
4.2km (2.6 m
per sec
Venus

Earth

Mars
b 700km 4,220 miles
4hr 37min 23sec
Mass 0.11 Earth

886.06 Earth days
Jupiter
Distance from Sun
778 300 000km
(483 600 000 miles)
143 000km (89 000 miles
Mass 318 Earth
Surface gravity
2 Earth
60.2km (37.4 miles

Saturn
(887 million miles)
Diameter (equatorial)
120 000km
(75 000 miles)
Rotation period
Mass 95 Earth
Earth
Escape velocity 36km
Sidereal period
29.46 Earth years

Uranus

Neptune
Distance from
mean 4 497 million km
(2 794 million miles)
Sidereal period 164.8 years
Pluto
mean 3940 m km
Sidereal period years

The Moon

The Moon is much the closest natural body to Earth in the sky. Its distance from the Earth is, on average, only 384,000km (239,000 miles), which is about equal to ten times the distance round the Earth's equator. It is a small world [1] with a diameter of 3,476km (2,160 miles), its mass is only 1/81 that of the Earth and the escape velocity is 2.4km (1.5 miles) per second, which is too low for the retention of an appreciable atmosphere.

Movements of the Moon

It is not entirely correct to say that the Moon revolves round the Earth. More properly, the Earth and Moon revolve round the "barycentre", or centre of gravity of the system. But because of the discrepancy between the masses of the two bodies, the barycentre lies well inside the terrestrial globe, so that the simple statement that "the Moon goes round the Earth" is good enough for most purposes. The revolution period is 27.3 days and this is also the time taken for the Moon to rotate once on its axis. As a result, the same hemisphere is always turned towards the Earth.

The Moon's path is not quite circular, so the apparent diameter of the disc varies within narrow limits. The familiar phases are due to the fact that the Moon does not always turn its daylight side towards the Earth [2]. The boundary between the day and night sides is known as the terminator; it is rough and jagged because the lunar surface is irregular. A peak will catch the rays of the rising Sun while the land below is still in darkness. Before the flight of the circumlunar probe Luna 3, in 1959, nothing was definitely known about the Moon's far side. Effects known as librations [3] (irregularities in the Moon's movement) produced by the Moon extend the visible area to a total of 59 per cent of the whole surface (although never more than 50 per cent at any one time).

Theories of origin

Although the Moon is officially ranked as the Earth's satellite, it seems disturbingly large to be a truly secondary body. There are other satellites in the Solar System that are larger than the Earth's moon (three members of Jupiter's family, one of Saturn's and one of Neptune's), yet all these move round giant planets. Thus Triton, the senior attendant of Neptune, has only 1/750 of the mass of its primary, although it is possibly larger than the planet Mercury and certainly larger than the Moon.

This being so, it may well be that the Earth–Moon system should more properly be regarded as a double planet, which leads to the problem of the Moon's origin. The tidal theory proposed by George Darwin (1845–1912) in the nineteenth century was popular for many years. According to this hypothesis the Earth and Moon were once a single body, rotating quickly and therefore becoming unstable. Eventually the globe became so distorted that part of it broke away and moved off to form the Moon.

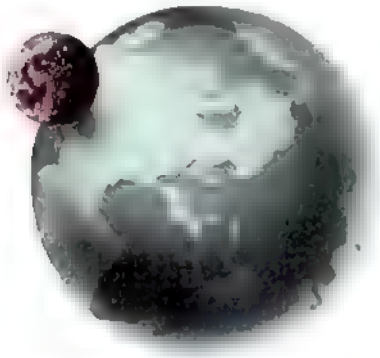
There are, however, mathematical objections to this theory which are so serious that few astronomers now support it in any form. It is much more likely that the Moon and the Earth were formed in the same way, from the solar nebula, either close together in space or else quite independently – in which case the Moon would have been 'captured' by the

CONNECTIONS

See also

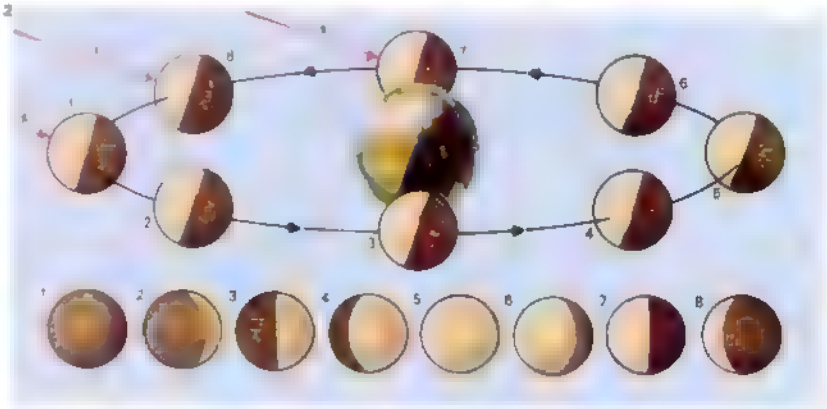
Moon panorama
The planet Earth

1



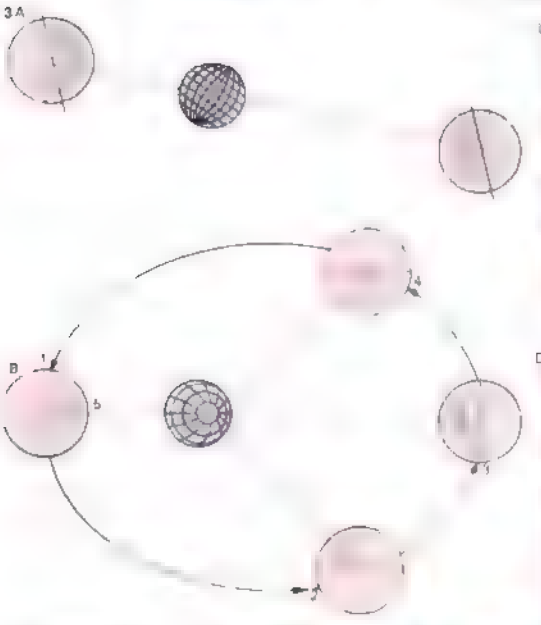
1 The Moon is a small world compared with the Earth. Its mass is much less and its specific gravity is lower. But the discrepancy between Earth and Moon is much less marked than with the satellites of other planets. With Neptune, for instance, the mass of its largest satellite, Triton, is only 1/750 of that of the planet. This is one reason why the Earth–Moon system may be regarded as a double planet.

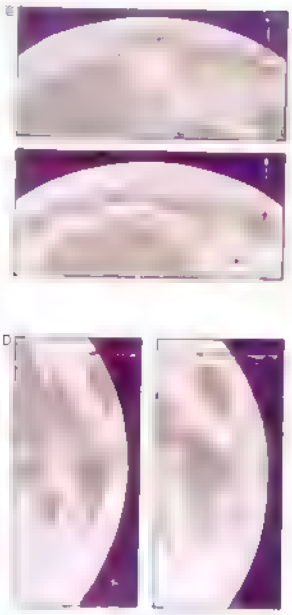
2



2 The phases of the Moon occur because the Moon has no light of its own. The daylight side reflects the Sun, the night side reflects "Earthshine". In the illustration, sunlight comes in from the upper left. In 1, the Moon's dark side is turned Earthwards (new) and the Moon cannot be seen – unless it passes directly in front of the Sun, producing a solar eclipse. Between 1 and 3 the Moon is crescent, at 3, half (first quarter), at 4, gibbous, and at 5, full. The Moon then wanes, through gibbous [6] to half or last quarter [7], crescent in the morning sky [8] and back to new.

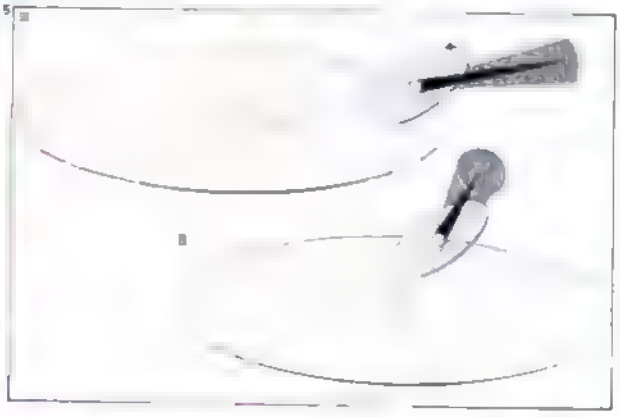
3 A



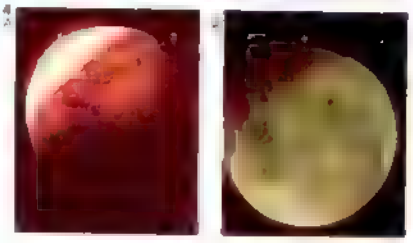


3 Librations are irregularities in the Moon's movement. Libration in latitude [A] occurs because the Moon's axis is tilted to its orbital plane, thus allowing views over the north [C] and south poles; libration in longitude [B] occurs when the Moon's speed of revolution changes slightly – it moves fastest at perigee [1] and slowest at apogee [3]. Its effect can be seen by tracing the position of point 5 through its locations in orbit from 1 to 4. This allows us to see a little farther round each limb (edge) [D].

5



4



4 During a total eclipse [A] the Moon does not vanish completely because a certain amount of light is refracted on to its surface by way of the Earth's atmosphere. The boundary between light and dark is never sharp. Partial eclipses [B] also occur. When the Moon passes through the area of penumbra that lies on either side of the main shadow cone, the visible effect becomes less striking.

5 Eclipses do not occur at every full Moon because the Moon's orbit is inclined at 5° to that of the Earth. At most full Moons the Moon passes either above or below the shadow [A]. A lunar eclipse [B] is produced when the Moon passes into the shadow cast by the Earth.

Earth later on. Based on our present state of knowledge, the former alternative seems to be the more plausible.

Surface features: seas and craters

The first telescopic maps of the Moon were drawn in 1609. Priority may well belong to Thomas Harriot (1560–1621), who drew a chart that shows many features in recognizable form. A longer and more systematic study was carried out from 1610 by Galileo, who described the mountains, the craters and the grey plains in some detail. The grey areas were named "seas" and the nomenclature has not been altered, although for centuries it has been known that there is no water in them. The names are usually given in Latin, thus the Sea of Clouds is Mare Nubium and the Ocean of Storms Oceanus Procellarum.

The walled structures commonly known as craters dominate the entire lunar scene. In size, they range from vast enclosures more than 240km (150 miles) in diameter down to tiny pits too small to be seen from Earth. A typical crater has a rampart that rises to only a modest height above the outer terrain, the floor is sunken, and there may be a central mountain or mountain group. In some the wall may be more than 3,000m (10,000ft) above the deepest part of the floor. There have been endless arguments about the origin of the craters. The main controversy centres on one point: were the craters produced by external forces (meteoritic impact) or by internal ones (vulcanism)? No doubt, like the Earth, the Moon has both types.

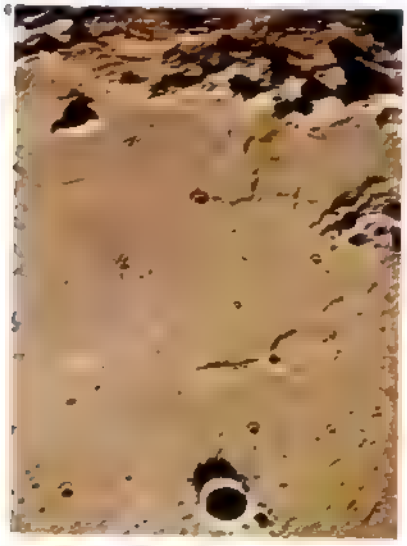
Some of the waterless seas, like the regular craters, are more or less circular with mountainous borders. The huge Mare Imbrium or Sea of Showers, for example, is bounded by the Apennines, Carpathians and Alps, although the mountain boundary is not continuous and there are wide gaps. The Apennines [7] are the most spectacular of the ranges; their loftiest peaks reach heights of more than 4,570m (15,000ft).

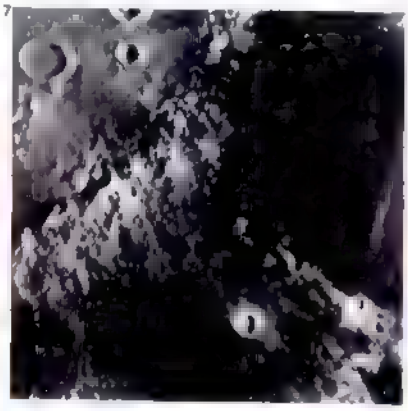
Other lunar features include hills, domes, with gentle slopes and often a summit craterlet or craterlets; occasional faults, and many crack-like features called clefts or rilles. Some of these markings can be identified even with the naked eye.

KEY

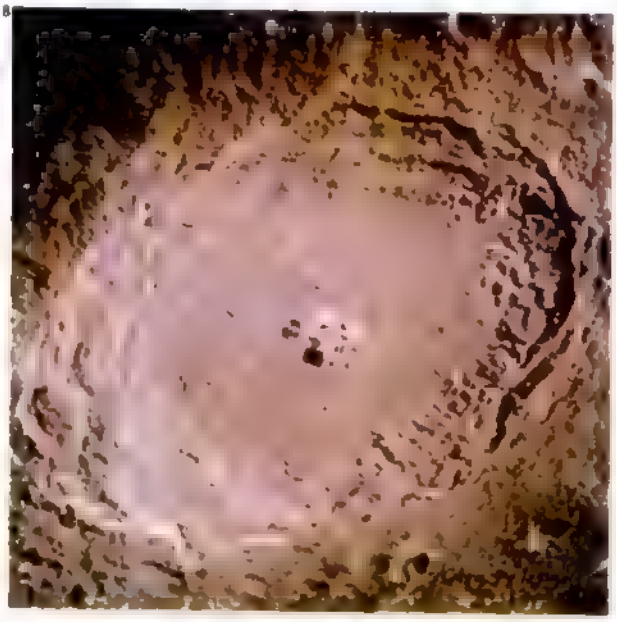
The dark "seas" and bright uplands of the full Moon are seen here. The crater Tycho dominates the southern hemisphere.

6 The Sinus Medii or Central Bay photographed from Apollo 10, is one of the relatively smooth mare areas.



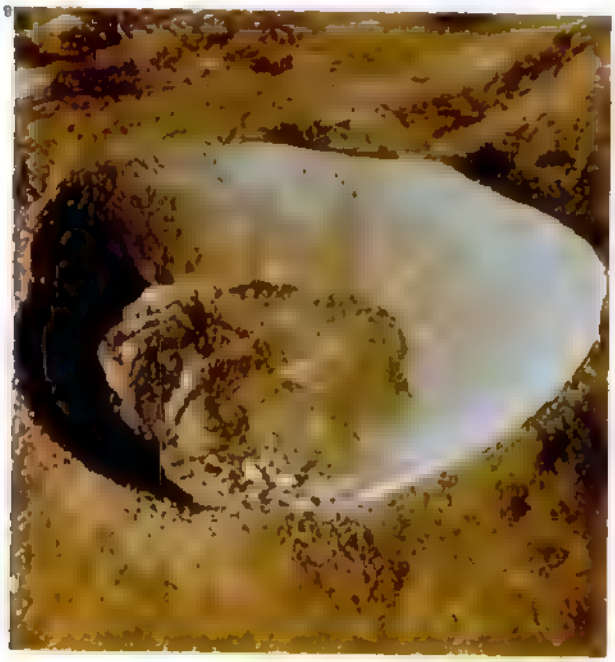


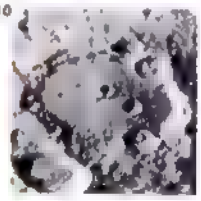
7 The lunar Apennines are here photographed with a 12in (30cm) reflector. The highest mountains rise to 4,570m (15,000ft) above the plains. They are by far the most spectacular of the lunar peaks.



8 The great crater Langrenus has massive terraced walls and a complicated central mountain group.

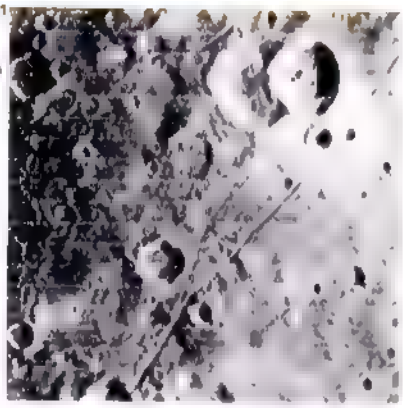
9 This small crater on the far side of the Moon was photographed by Apollo 10.

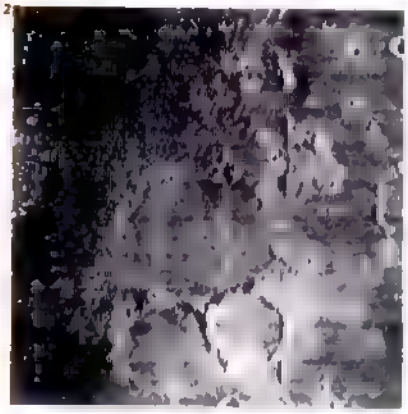




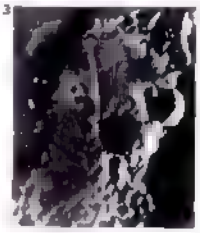
10 Clavius, photographed from Earth is a walled plain 230 km (143 miles) wide with superimposed craters.

11 Part of the Sirsalis Rille, a telescopically visible collapse feature that is 32km (20 miles) in diameter.





12 An Orbiter photograph of the connecting walls of three plains – Fra Mauro, Bonpland and Parry – is a typical example of damaged outer walls.



13 The Straight Wall, the Moon's best known fault, is illuminated here by sunlight.

Moon missions

Flight to the Moon became a practical possibility after the opening of the Space Age in October 1957 when the USSR launched Sputnik 1, a satellite that circled the Earth. Two years later, she sent three vehicles on lunar missions. The first of these (Luna or Lunik 1) bypassed the Moon and sent back useful information, notably that there was no appreciable magnetic field. Luna 2 crash-landed on the Mare Imbrium in September 1959 and in October Luna 3 [1] went on a round trip sending back the first photographs of the Moon's far side. This proved to be just as mountainous, crater-scarred and sterile as the side seen from Earth, but there were no comparable seas.

Early unmanned explorations

The American Ranger programme [2] introduced a new phase in man's knowledge of the Moon. These Rangers were designed to send back close-range photographs before crashing onto the Moon. Ranger 7, in 1964, was the first successful probe of this kind. Two more followed, the last of them came down in the prominent crater Alphonsus near the centre of the Moon's disc as seen from the Earth.

In January 1966 the USSR achieved a major triumph by soft-landing an automatic probe, Luna 9, on the surface of the Moon. Its cameras showed a landscape that looked remarkably like a lava plain, with hummocks and crater pits everywhere. Luna 9 was of special importance because it finally disproved a curious theory that the lunar seas were filled with soft, treacherous dust to a depth of several hundred metres. The landing showed that the Moon's surface layer, or regolith, is firm enough to support the weight of a spacecraft [13].

Lunar mapping was more or less completed during the two years following August 1966. Five American Orbiters [4] moved around the Moon in closed paths, sending back amazingly detailed pictures. The Americans also soft-landed several vehicles – the Surveyors. On 17 January 1968, Surveyor 7 came down near the crater Tycho, in the southern uplands, and sent back high-quality pictures of the outer slopes of the wall of the crater.

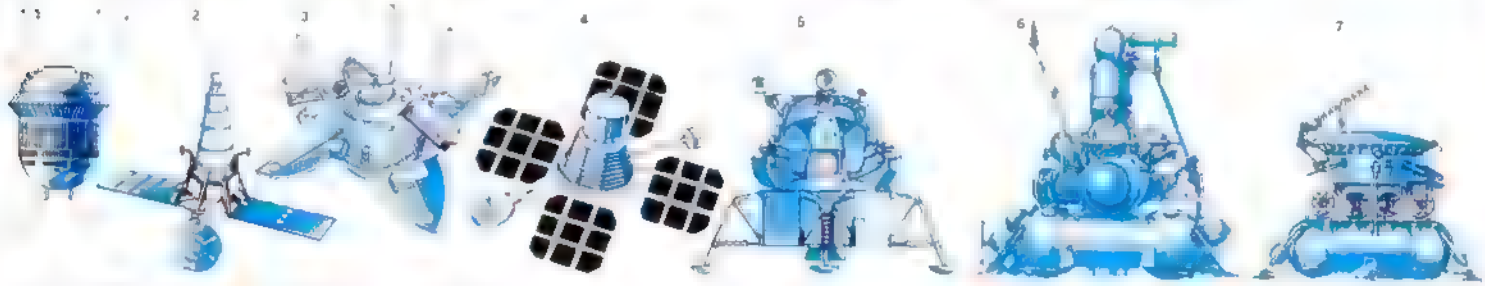
The USSR continued its programme of unmanned exploration into the 1970s. It achieved a major triumph with Luna 16 [6] in 1970, which landed in the Mare Foecunditatis and then returned, bringing samples of lunar rocks. Later in the same year Luna 17 landed in the Mare Imbrium. From it emerged Lunokhod 1 [7], an eight-wheeled craft powered by solar batteries and guided from Earth. After extensive exploration it ceased to function on 4 October 1971. Lunokhod 2 followed in 1973.

Men on the Moon

From the mid-1960s onwards the Americans had been concentrating on the Apollo programme of sending men to the Moon. This culminated in July 1969, when Neil Armstrong (1930–) and Edwin Aldrin (1930–) left Eagle [5], the lunar module of Apollo 11 [8], and made the historic "one small step" onto the lunar surface. After collecting samples of lunar material [14] and leaving recording instruments behind [12], the two astronauts returned to their module and rejoined the third member of the expedi-

CONNECTIONS

See also



1 Luna 3 was the probe that made the first circumlunar flight (October 1959). It sent back pictures of the far side.

2 Ranger 8 (1965) crash landed near the crater Del ambre, but sent back excellent pictures before impact.

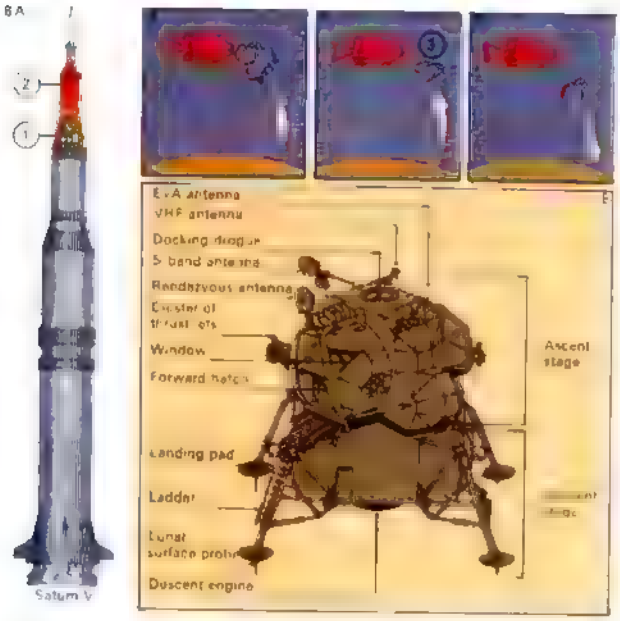
3 Luna 13 made the first successful soft landing on 21 December 1966 and sent back pictures from the surface.

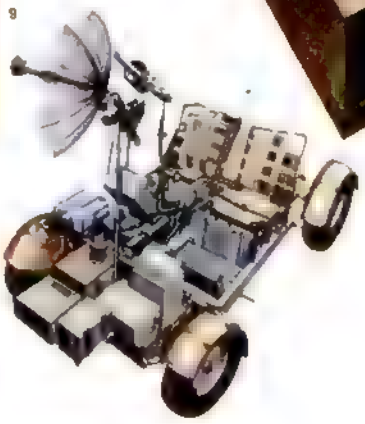
4 Orbiter 5 (1968) was the last of the series of Orbiters, which have provided full photographic surface coverage.

5 Eagle (1969), the lunar module of Apollo 11, took Armstrong and Aldrin to Tranquility Base on the Moon.

6 Luna 16 (1970) landed in the Mare Foecunditatis and collected samples of lunar material to bring back.

7 Lunokhod 1, which landed with Luna 17 in November 1970, took photographs and crawled about the surface collecting data.





8 The Apollo Moon programme used a Saturn rocket (A) to carry the lunar module (1) and the command and service modules (2). The lunar module, holding two astronauts, descended to the Moon (B). Its upper part (3) later left, blasting off from the descent stage back into orbit (C) to rendezvous with the third member of the team, still in orbit round the Moon. The module was then jettisoned (D) and crashed onto the Moon. The lunar module is shown in detail (E).

9 The Lunar Roving Vehicle (LRV) was carried by the last three Apollos (15, 16 and 17) and was used by the astronauts to drive across the Moon's surface for considerable distances. Its speed ranges from 8 to 16km/h (5–10mph).

10 The ALSEP of Apollo 17 has a central station and thermal generator (1, 2) to provide the main power. The atmospheric composition experiment (3) analyses any residual lunar atmosphere. The ejecta and meteorite experiment (4) detects any impact from meteorite bodies and the gravimeter (5) measures any gravitational anomalies. The geophone experiment (6) involves artificial explosions that will help in studies of the Moon's surface layers. There are also seismic measuring devices (7) and core sampling equipment (8). The solar wind experiment (9) was carried out on earlier Apollo missions.

tion, Michael Collins (1930–), who was orbiting the Moon in the command module

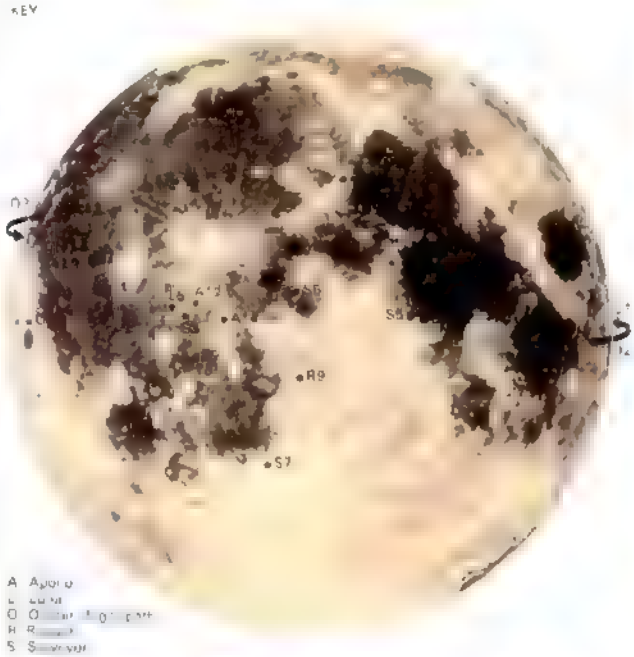
Apollo 12 followed later in 1969; the explorers, astronauts Charles Conrad (1930–) and Alan Bean (1932–), landed near a previous automatic probe, Surveyor 3, and were able to bring parts of it home. With the next mission, Apollo 13, came the first real failure. An explosion in the service module of the spacecraft during the outward journey put the main power supplies out of action. The lunar landing was cancelled and it was only by a combination of courage, skill and luck on the part of both the astronauts and the operators at mission control that tragedy was averted. Since then there have been four more Apollo landings. The series ended in 1972 with Apollo 17, which landed in the Taurus–Littrow region, manned by Eugene Cernan (1934–) and Harrison Schmitt (1935–), a professional geologist

Each successful mission has deployed what is called an Apollo Lunar Surface Experimental Package (ALSEP) [10]. Various investigations have been carried out and our knowledge of the Moon has improved immensely, even though astronomers still argue about the origin of the main craters. In 1979, most ALSEPSs, were still capable of transmitting information to Earth, although they have been switched off

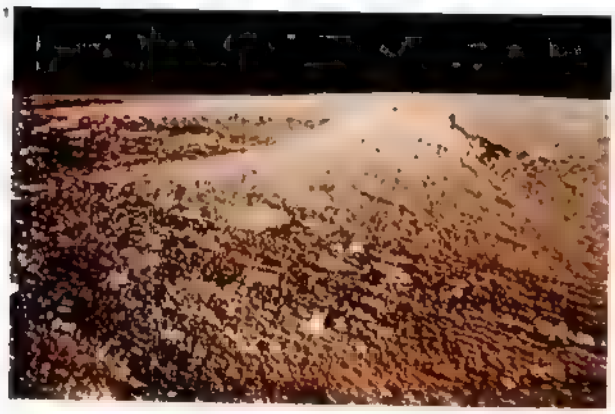
Lunar landscape

Conditions on the Moon are unfamiliar. An astronaut has only one-sixth of his normal weight although his mass is unaltered. There is virtually no local surface colour and the lunar sky is black even when the Sun is above the horizon. The lunar day is long, because the Moon spins slowly. So far all the landings have been made in the early morning on a selected region of the Moon

The Moon is not a welcoming world. The temperatures range between about 90°C (195°F) at noon on the equator down to well below −130°C (−200°F) at night. There is no air or water and we are now sure that there has never been any life. Yet the Moon is of tremendous importance to man. Before the year 2000, a permanent lunar base may well have been established there for scientific and astronomical research



11

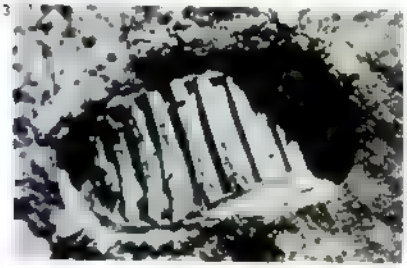


11 This lunar landscape photograph, taken on the Apollo 11 mission, illustrates several surface details encountered by the astronauts, although the area shown is smooth by lunar standards

12

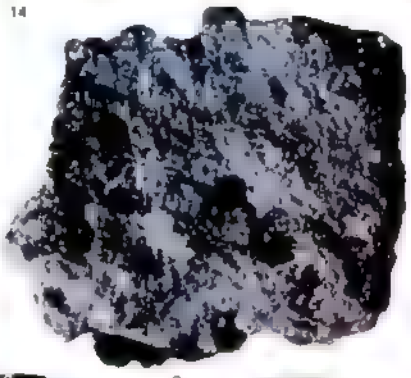
12 During the Apollo 11 mission Armstrong and Aldrin set up a lunar seismometer to measure ground tremors. The seismometer is similar to instruments used on Earth but can be more sensitive because the Moon is seismically "quiet". The particular one failed after a short period but similar instruments left on the Moon by later Apollos have shown that minor tremors do occur frequently

13



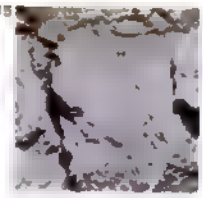
13 Edwin Aldrin's footprint on the lunar surface had a depth of penetration of less than 2.5cm (1in)

14



14 Rock samples brought back from the Moon by Apollo 11 included this one, of basaltic type. It shows virtually no evidence of any hydrated material. Samples brought back from the later Apollo and Luna missions show that there are numerous types of lunar rock, most of the minerals of which they are composed occur on Earth

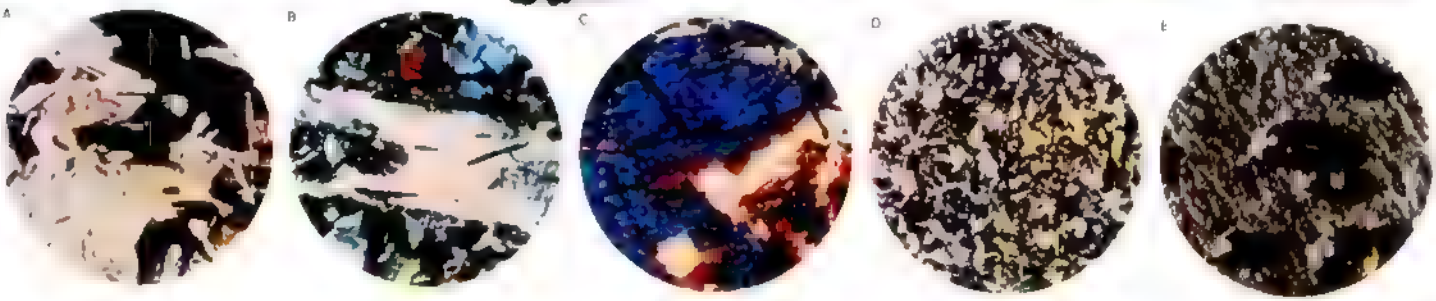
15



15 A section of the rock in illustration 14 is enlarged here, the fine structure is clearly shown. This was one of the first rocks to be analysed

16 Microstructure photographs (A–E) of samples brought back by Armstrong and Aldrin from the Mare Tranquillitatis are shown here. Substances identified include plagioclase, ilmenite, pyroxene and microscopic rubies. There are many "glass marbles", well shown in E. These particles are small, the largest of them being only 0.5mm (0.02in) in diameter

16 A B C D E



The Moon's structure

Analysis of the samples brought back from the Moon by the American Apollo missions and the Soviet unmanned probes have established that the Earth and the Moon are about the same age (between 4,500 and 5,000 million years, or 4.5 and 5 aeons). But because the masses of the Earth and the Moon are so different they have undergone different evolutionary sequences [1]

The surface of the Moon

The nature of the Moon's surface is intimately bound up with the problem of the origin of its craters and other features and this has led to endless arguments that have not been resolved even by the Apollo results Some strange theories of crater origin have been advanced (ranging from coral atolls to atomic bombs), but the whole problem centres on whether the craters were formed by internal action or by external bombardment The first of these rival theories is usually called the "volcanic" theory [3], the second is known as the "meteoric" or "impact" theory [2]

Both processes must, in fact, have operated to some extent for both kinds of craters occur on Earth, and no doubt do so on the Moon as well What must be decided is which of the two processes played the more important role. Opinion is sharply divided, although some of the features, such as the small chain craters, are undeniably "volcanic" in the broad sense.

Many efforts have been made to link the main lunar craters with terrestrial impact craters such as that at Arizona, USA, although the scales are different; if transferred to the Moon, the Arizona crater would appear insignificant. On the other hand, supporters of the volcanic theory point out that the distribution of the Moon's craters is not random [4]; for instance, the great crater plains tend to appear in distinct lines. When one crater formation breaks into another the smaller crater almost always intrudes into the larger. This is easier to explain with a theory of internal origin than with one of impact. It is also true that the lunar rock samples are essentially volcanic, although often affected by impact, and that the incidence of meteoritic material appears to be comparatively scarce on the Moon's surface.

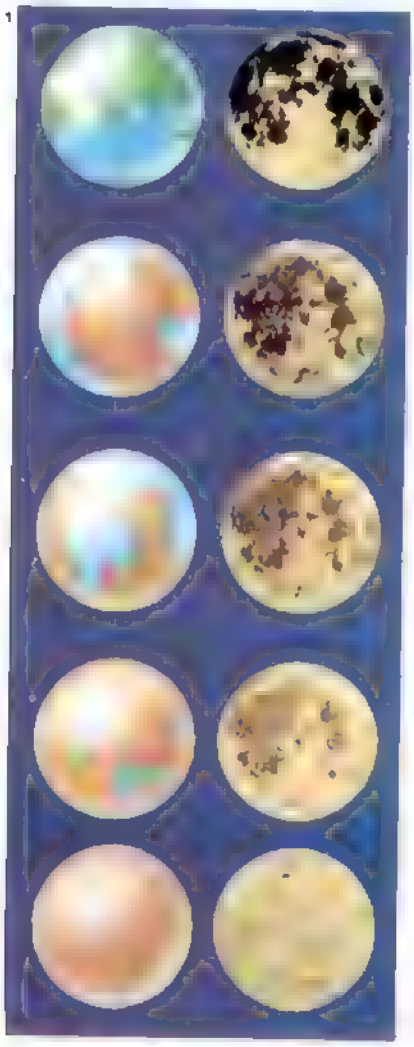
It is also questionable whether the circular maria, or lunar "seas", are essentially similar to the large craters in origin. At least there is some reliable information about their age and it seems probable that the chief maria (Imbrium, Serenitatis, Crisium and the rest) were formed about 4,000 million years ago the Mare Orientale is probably the youngest of them with an age reliably estimated at about 3 8 aeons.

The filling of the lunar "seas"

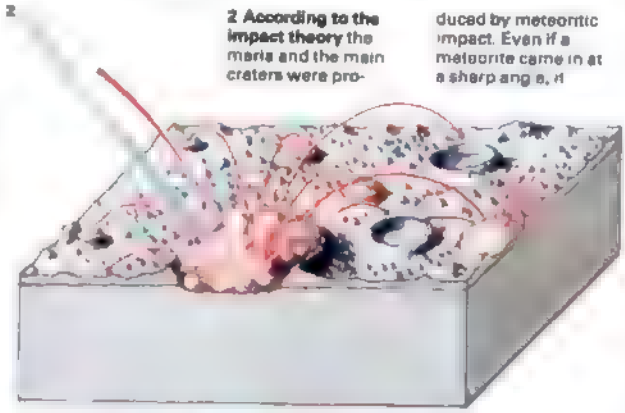
Most experts agree that when the mare basins were formed they did not contain lava; the question of whether they were created by internal (endogenic) or external (exogenic) factors makes no difference to this conclusion. Between 3 8 and 3.2 aeons ago, the mare basins began to be filled by lava that poured out from beneath the crust and finally produced the aspect that is seen today Because the eruptions were not confined during about a million years, the mare surfaces, although apparently simple, form a patchwork of overlapping lava flows [7]

CONNECTIONS

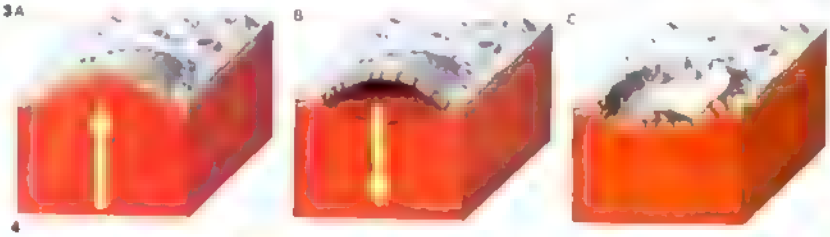
See also



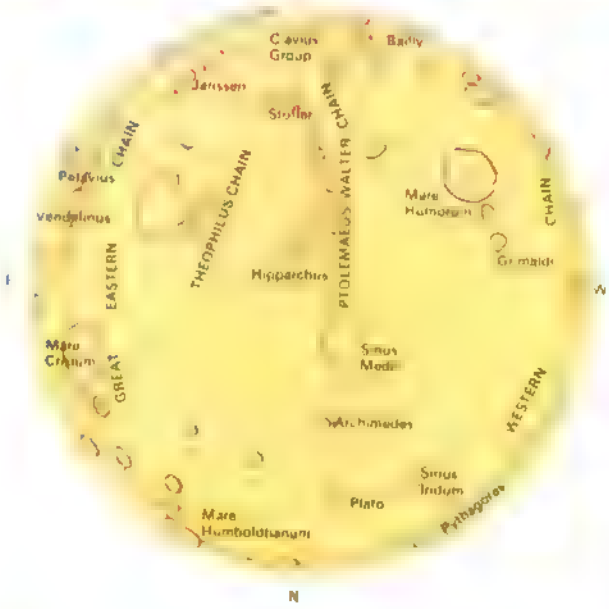
1 The mare basins were formed, either by internal accretion or by impact, at an early stage in the history of the development of Moon and Earth surfaces (4,000 million years ago). The general aspect of the Moon then must not have been dissimilar to that of today, although the basins were not filled. Little is known about surface details on Earth at this time. The surface of both the Earth and the Moon then remained the same for a considerable period. Some 2,000 million years ago the basins on the Moon were filled in, 1,000 million years after that, lunar activity was at an end. Geological techniques give some idea of the appearance of the Earth at that time, since when it has developed markedly into its present-day form.



2 According to the impact theory the maria and the main craters were produced by meteoritic impact. Even if a meteorite came in at a sharp angle, it would still produce a circular formation This theory is a popular one today



3 The rival volcanic theory states that when the lunar surface was hot and plastic (A), domes were produced, by magmatic convection for example. On cooling the underlying material sank (B), leaving a void The surface layer collapsed, forming a crater (C) Central peaks are caused by penetration of magma.



4 The distribution of the lunar craters and walled formations is clearly crucial to any considerations of their origin. The small features tend to line up in chains and many of the so-called rilles are in part crater chains, there is no serious doubt that these are of internal origin. On the Earth turned hemisphere the major formations also tend to line up There are important chains, such as those including Vendelinus and Petavius (and also the Mare Crisium) in the east and the Grimaldi chain in the west It has been argued that the important features formed along lines of crustal weakness produced by the gravitational influence of the Earth

A certain amount of cratering occurred during this period. The ray-craters such as Tycho and Copernicus are probably the youngest of the major features and the age of Copernicus may be even less than one aeon. Subsequently the main activity ceased and only small (mainly impact) craters have been formed since that time.

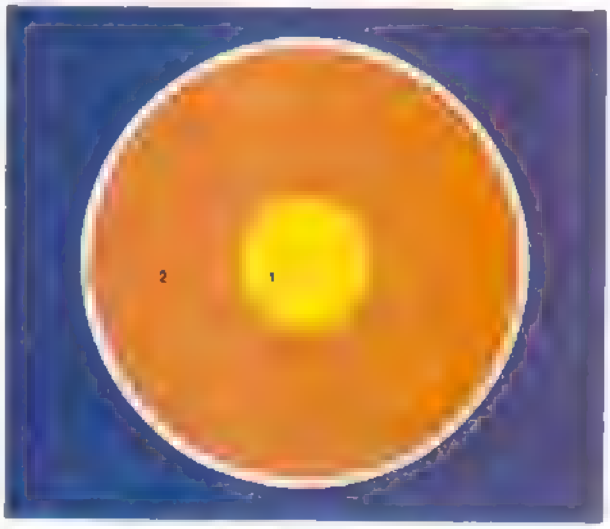
Evolution of the Moon

Curiously, more information exists about the geological evolution of the Moon than that of the Earth. Unlike the Earth, which has a long history of continual erosion, the Moon has not suffered erosion for a long time. Two thousand million years ago the Moon may have looked much the same as it does today, whereas the Earth would perhaps have been completely unrecognizable.

The Apollo seismometers have been able to record "moonquakes" [5] and there is now no doubt that a certain amount of volcanic activity lingers on. Some of the moonquakes occur close to the crust, but others are deep-seated – up to half-way to the Moon's core [6]. It has also been established that the Moon may have a hot core, so that the old idea of a globe that is cold throughout may be incorrect [Key]. Studies of moonquake records indicate that if a molten core exists it must be smaller than that of the Earth, both relatively and absolutely. Above this is the so-called asthenosphere, or zone of partial melting; above that the thick mantle, topped by a crust; and finally the rubbly regolith, which has a depth of 100m (328ft). There is virtually no general lunar magnetic field now, although some areas are locally magnetized; it seems that in the remote past the Moon had an appreciable general field that generally weakened and has now disappeared.

Earth-based lunar observers have recorded various minor "events" that may indicate the emission of gas from below the crust. These are known commonly as TLP, or Transient Lunar Phenomena [8]. They are thought to be commonest near perigee, when the Moon is at its closest to the Earth. The Earth's gravitational pull then produces the greatest strain on the lunar rocks and there may be a link between this and the positions of the epicentres of moonquakes.

KEY

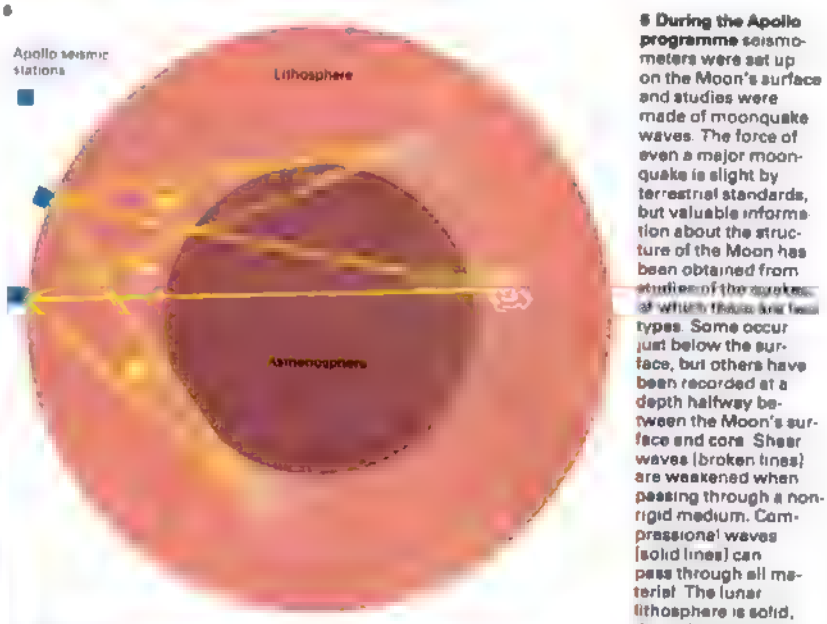


The core of the Moon [1] is believed to be much smaller than that of the Earth, both relatively and absolutely, and it is probably extremely rich in iron. Surrounding it is the lunar mantle [2]. This is overlaid by the crust [3], which is covered by a rocky "topsoil" (regolith).

6

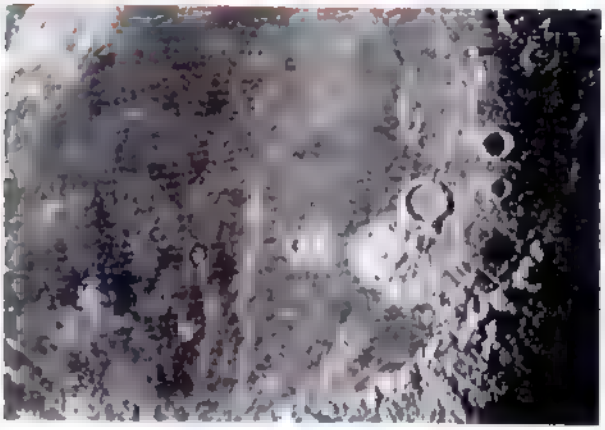
5 The lunar maria were formed when floods of lava erupted onto the Moon's surface, filling in basins previously excavated by planetesimal impact. They did not erupt in one burst of volcanic activity, but over a period of about 1,000 million years. The oldest basalts are thought to have been produced at a depth of 150km (95 miles) [1]. The more recent rocks were generated later at a depth of about 240km (150 miles) [2].

6



6 During the Apollo programme seismometers were set up on the Moon's surface and studies were made of moonquake waves. The force of even a major moonquake is slight by terrestrial standards, but valuable information about the structure of the Moon has been obtained from studies of the quakes, of which there are two types. Some occur just below the surface, but others have been recorded at a depth halfway between the Moon's surface and core. Shear waves (broken lines) are weakened when passing through a non-rigid medium. Compressional waves (solid lines) can pass through all material. The lunar lithosphere is solid, the asthenosphere is a near-melting zone.

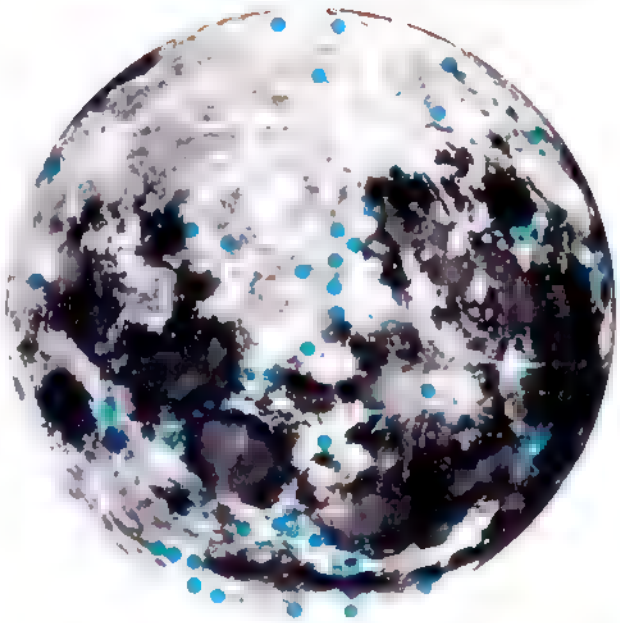
7 A



B

7 In the Mare Imbrium (A) the darkest red coloration on the map (B) represents the youngest of the lava flows, which is the area least affected by cratering. The lightest red represents the oldest flow. The most extensive flow dates back approximately 3.3 thousand million years.

8



8 Transient Lunar Phenomena (TLP) do exist on the Moon, as has been established in recent years. The TLP-prone sites are not distributed at random: they tend to congregate round the borders of the circular maria and in regions rich in rilles. The most active area on the Moon is that of the brilliant crater Aristarchus.

Moon maps

Even with the naked eye the Moon can be seen in considerable detail and binoculars or any telescope provide the observer with a seemingly inexhaustible panorama. Obviously the view depends on the angle of sunlight over the area being studied. A crater is at its most prominent when near the terminator (the boundary between the day and night hemispheres of the Moon), so that its floor is wholly or partially shadow-filled. Even a major crater may become difficult to identify under a high angle of illumination, unless its floor is either particularly dark or particularly brilliant. In general, the lunar surface has a low reflecting power, or albedo, of about seven per cent – that is to say it reflects only about seven per cent of the sunlight falling on it – but the brightest craters possess walls and central peaks that have albedos of more than 15 per cent.

The northern hemisphere

The northern hemisphere of the Earth-turned side of the Moon is dominated by two great seas, the Mare Imbrium (Sea of Showers) and the Mare Serenitatis (Sea of Serenity), both of which are approximately circular, although foreshortening makes them appear somewhat elliptical. The Mare Imbrium has mountainous borders for much of its outline, including the majestic Apennines with peaks rising to some 4,570m (15,000ft). Between the Apennines and the rather lower Caucasus Mountains there is a gap linking the Mare Imbrium to the Mare Serenitatis. The 95km (60 mile) dark-floored crater Plato lies in the region of the Alps and also in this range is the remarkable Alpine Valley, 130km (80 miles) long.

On the floor of the Mare Imbrium there are several major craters, including the 80km (50 mile) Archimedes and its two smaller but deeper companions called Aristillus and Autolycus. The Mare Serenitatis includes no crater of this size; the largest, Bessel, is only 39km (24 miles) across.

The Mare Tranquillitatis (Sea of Tranquillity), adjoining the Mare Serenitatis to the south, is less regular in form and is presumably older. It was in this sea that two of the astronauts of Apollo 11 made the historic pioneer lunar landing in July 1969.

The Mare Crisium (Sea of Crises), not far from the limb, is smaller but well marked and is easily visible with the naked eye. Of the other seas in this hemisphere the largest is the Oceanus Procellarum (Ocean of Storms), separated from the Mare Imbrium by the relatively modest Carpathian Mountains. Aristarchus, on the Oceanus Procellarum, is generally regarded as the brightest crater on the Moon and can often be seen shining prominently even when on the dark side of the terminator, lit only by light reflected from the Earth. The crater Copernicus, south of the Carpathian Mountains, is the centre of a bright ray system.

Another interesting feature in the northern hemisphere is a great bay, the Sinus Iridum (Bay of Rainbows), which leads off the Mare Imbrium. When the Sun is rising over it, illuminating the mountainous border, the effect is comparable to a jewelled handle.

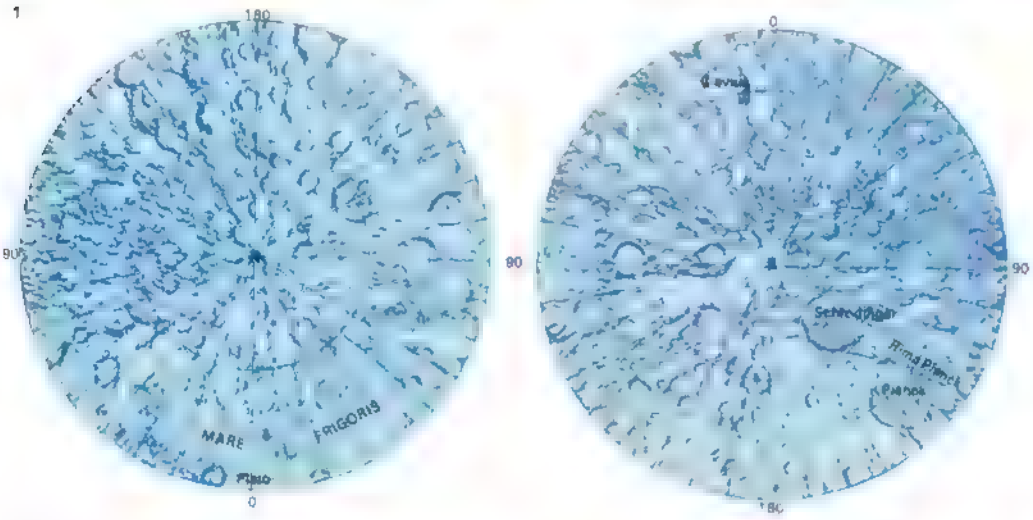
The southern hemisphere

Slightly south of the equator are the great plains of which Ptolemaeus is the senior member. Ptolemaeus is almost 160km (100

CONNECTIONS

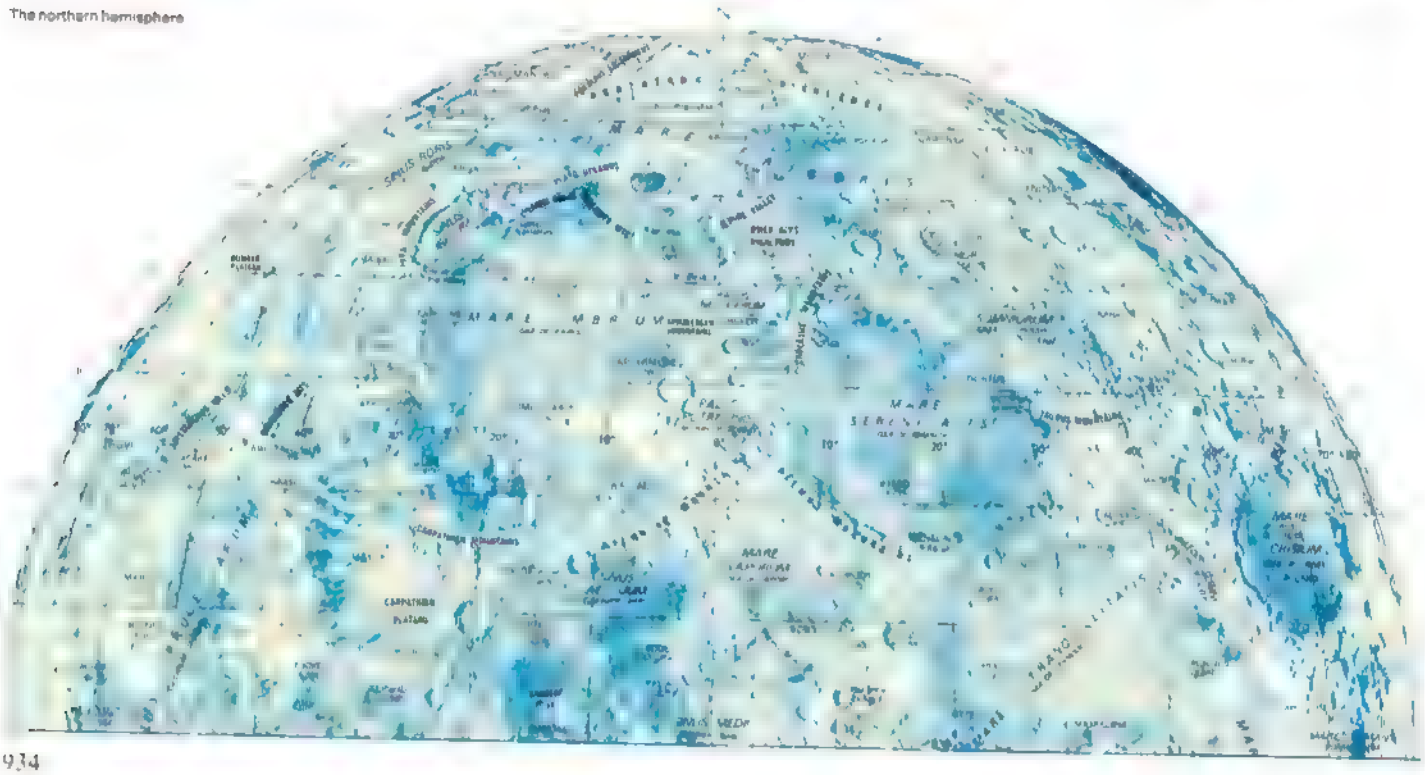
See also

[illegible]
Moon [illegible]
The Moon [illegible]
[illegible]
Moon [illegible]



1 The polar zones of the Moon are difficult to study from Earth because of extreme foreshortening. Some of the areas in these maps cannot be seen at all and our knowledge of them is derived mainly from photographs sent back by the five highly successful Orbiters. Mapping of part of the south polar region is not yet quite complete, although future probes will take new photographs of the region. Both polar zones are heavily cratered and there are many large walled plains. On the northern chart Mare Frigoris, visible from the Earth, is shown, Plato appears near the bottom of the map. One of the interesting features in the south polar map is the large walled plain Schrödinger, on the far hemisphere, together with a similar structure, Planck. Associated with these two features is a long rille, the Rima Planck (also called the Schrödinger Valley). The area in the upper section of the south polar chart can be studied from Earth, the great walled plain Clavius is clearly visible here.

The northern hemisphere



miles) in diameter, with a relatively level darkish floor. Adjoining it is the rather smaller Alphonsus, with a central mountain group and a system of rilles on its floor. In 1958, from Alphonsus the Russian astronomer N. A. Kozyrev recorded a reddish glow – one of the best authenticated examples of a Transient Lunar Phenomenon (TLP). This, he believed, indicated a certain amount of surface or sub-surface activity which he interpreted as being of volcanic origin. The third member of the Ptolemaeus chain, Arzachel, is smaller than Alphonsus, but deeper, with a higher central peak.

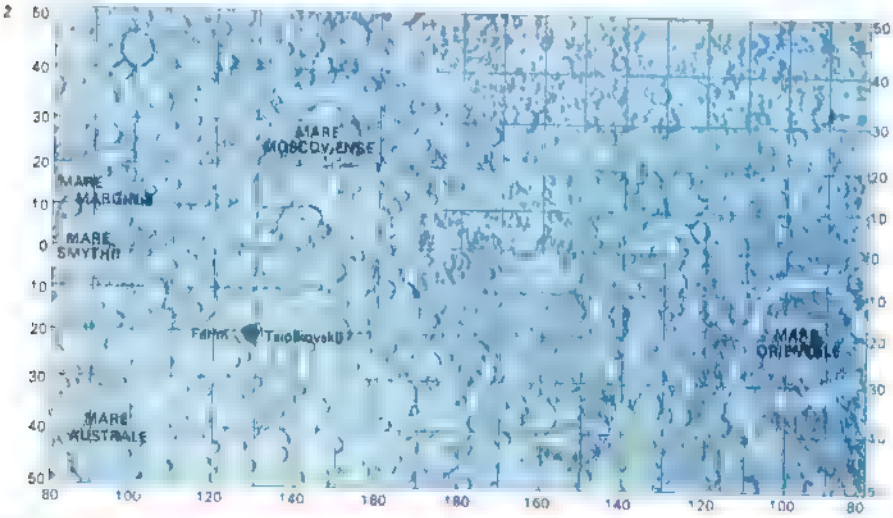
The southern part of the Moon consists largely of rugged upland, although there are some sea areas – the Mare Nubium (Sea of Clouds) and the smaller Mare Humorum (Sea of Moisture). On the former, not far from Arzachel, is the Rupes Recta (Straight Wall), which is a major fault in the surface, with a length of 130km (80 miles) and with a height between the crest and the bottom of 240m (800ft).

Among other major walled plains are the dark-floored Schickard and the 230km (144 mile) Clavius, which has a chain of craters inside it. North of the Southern Highlands is the crater Tycho, called the "metropolitan crater" of the Moon because of its system of bright rays, which are brilliant and extensive. Near full Moon, as seen from Earth, they dominate the entire southern hemisphere and render virtually invisible even large walled plains in the vicinity. Tycho is 86km (54 miles) in diameter, with massive walls, and even under low illumination when the rays are not properly visible, it remains one of the most prominent craters on the Moon.

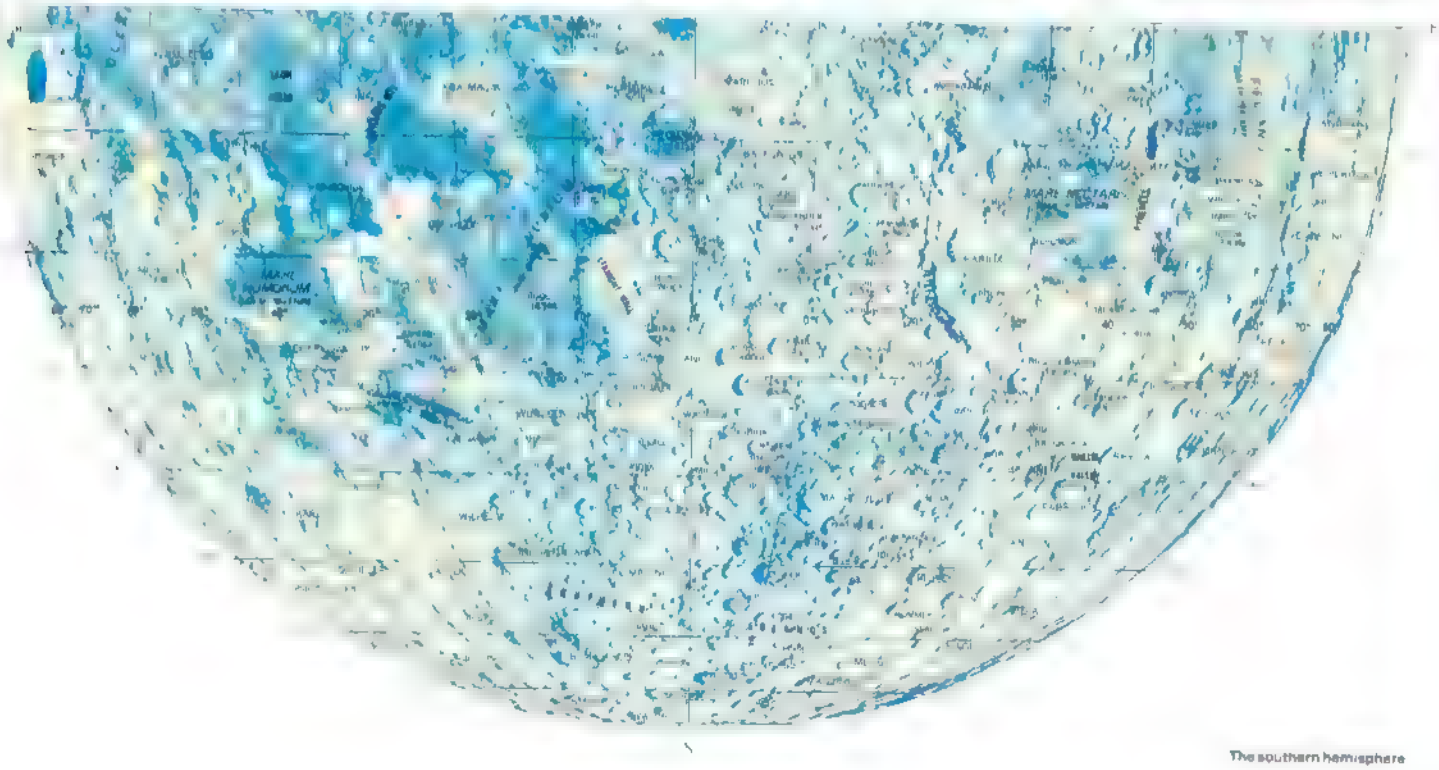
The far side of the Moon

The libration regions of the Moon were not well mapped until the age of space probes. Today we also have full information about the Moon's far side, although it has been seen "direct" only by the Apollo astronauts who have been round the Moon. Here there are no major maria, but there are walled structures of all kinds; of special interest is Tsiolkovskii, which has a very dark floor and was identifiable on the photographs sent back by Luna 3 in October 1959.

This Apollo 16 photograph shows parts of both the Earth-turned hemisphere and the far side of the Moon.



2 A map of the far side of the Moon shows no major maria similar to those on the familiar side of the Moon. The Mare Orientale is visible from Earth, although very foreshortened. On the far side there are craters of all kinds and the whole surface is "crowded uplands". The most striking feature is the dark-floored Tsiolkovskii, first seen on the Luna 3 photographs in 1959. It abuts onto another structure Fermi, which has a light floor. On the Luna 3 photographs a bright streak showed up, thought to be a major mountain chain, the Russians named it the "Soviet Range", but subsequently it was found to be nothing more than a bright ray. The Mare Moscoviense (Moscow Sea) was also identified on the Luna 3 pictures. The question of naming the features on the far side caused some controversy. The nomenclature given here is that now officially adopted by a special committee set up by the International Astronomical Union. The Mare Marginis, Mare Smythii and Mare Australe may also be seen from the Earth.



The southern hemisphere

Moon panorama

The Moon is a world of considerable variety. As well as the marked differences between the Earth-turned and the far hemispheres, there are also obvious changes of scenery on the familiar face of the Moon. For instance, the southwestern quadrant is dominated by rough uplands, with large craters and walled plains, whereas the northeastern quadrant contains vast stretches of mare surface.

In particular, there is the region of Aristarchus [3], the brightest crater on the Moon, which is of special interest because so many local obscurations have been reported there by Earth-based observers. William Herschel (1738–1822), possibly the greatest practical observer of all time, on several occasions mistook Aristarchus for a volcano in eruption when he saw it shining conspicuously from that part of the Moon illuminated only by Earthshine.

Aristarchus is not the only feature in or near which activity has been suspected. Another is the walled plain Alphonsus [1], in the great Ptolemaeus chain near the centre of the Earth-turned hemisphere. Although Alphonsus and Aristarchus are so different, they do have one thing in common: they lie in regions rich in rilles or clefts. The same is true of most other regions with mild activity.

The Moon before Orbiter

Before the space probe era man's knowledge of the Moon was bound to be limited, despite the fact that the lack of a lunar atmosphere makes all the details more distinct. Measurements of the positions on the disc of various lunar features were possible and although work undertaken by S. A. Saunder and J. A. Hardcastle in 1907–9 was of value (the Saunder–Hardcastle measures are still referred to today) certain areas could still not be resolved. In particular, little was known about the limb regions, which are hopelessly foreshortened as seen from the Earth: when a crater is seen under such conditions it may be impossible to distinguish it from a ridge. Neither could there be any positive information about the Moon's far side. Research was carried out into the lunar rays that were seen to come from the far side, by "plotting back": it was possible to fix the positions of a few ray-centres with reasonable accuracy, but the distribution of the features on the far side remained unknown. It was, however, significant that none of the major maria on the Earth-turned side extends on to the far side – apart from the Mare Orientale, whose nature was not then realized.

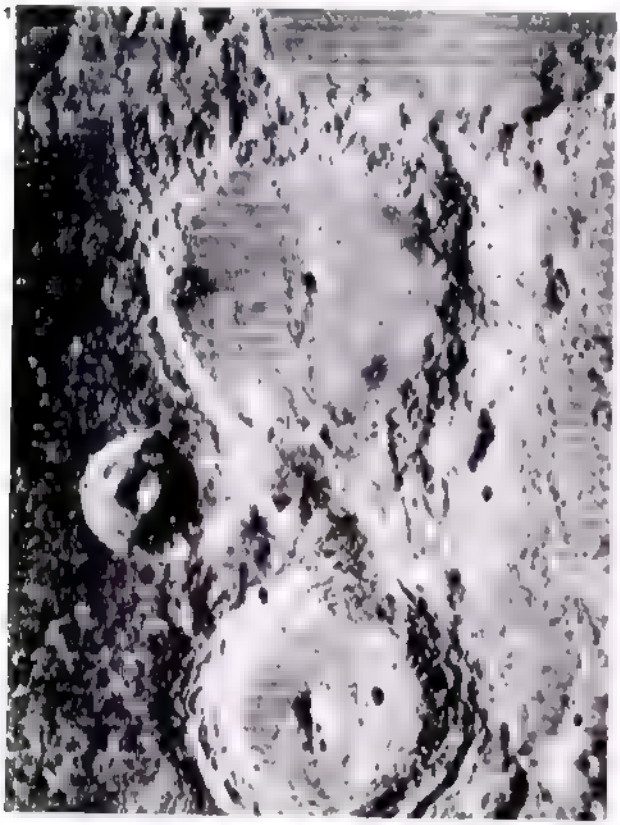
Photographic missions

The original photographs sent back by the Russian probe Luna 3 were of immense value, but by modern standards they were blurred and indistinct and thus widely misinterpreted [Key]. In particular, a long feature seen stretching across the disc was taken to be a major mountain range and was dignified by the title of the Soviet Mountains, but later photographs showed it to be nothing more than a bright ray. It was therefore not until the advent of the Orbiters that lunar research made its greatest stride forward, although the three successful Rangers, which crash-landed but managed to send back many thousands of useful pictures during the last minutes of their flights, also played a significant part.

Despite the comprehensive coverage of

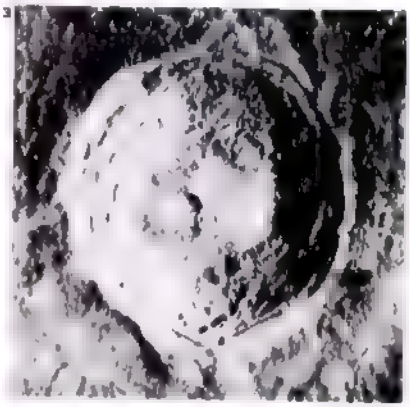
CONNECTIONS

See also



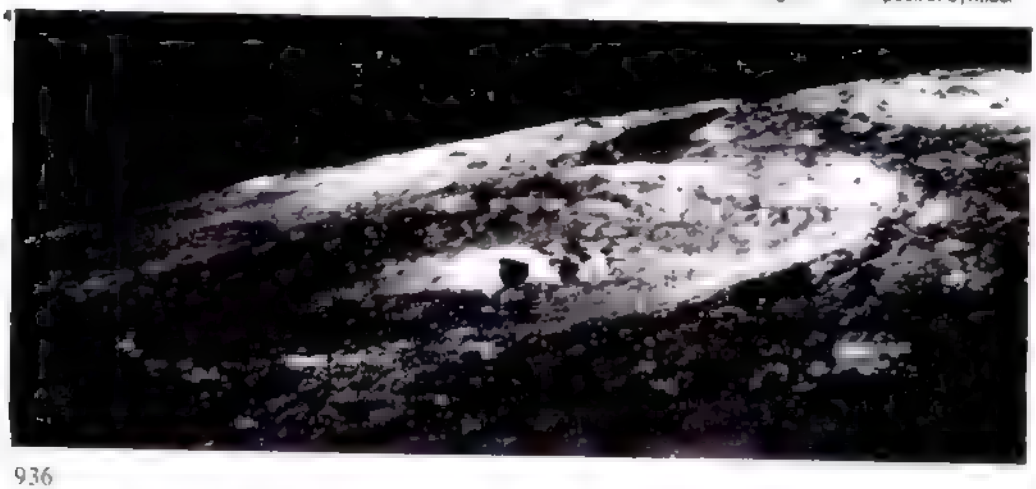
1 Alphonsus, the walled plain in which Ranger 9 landed and in which mild activity has been reported, is a member of the Ptolemaeus chain. Ptolemaeus itself is to the north (above) and Arzachel to the south.

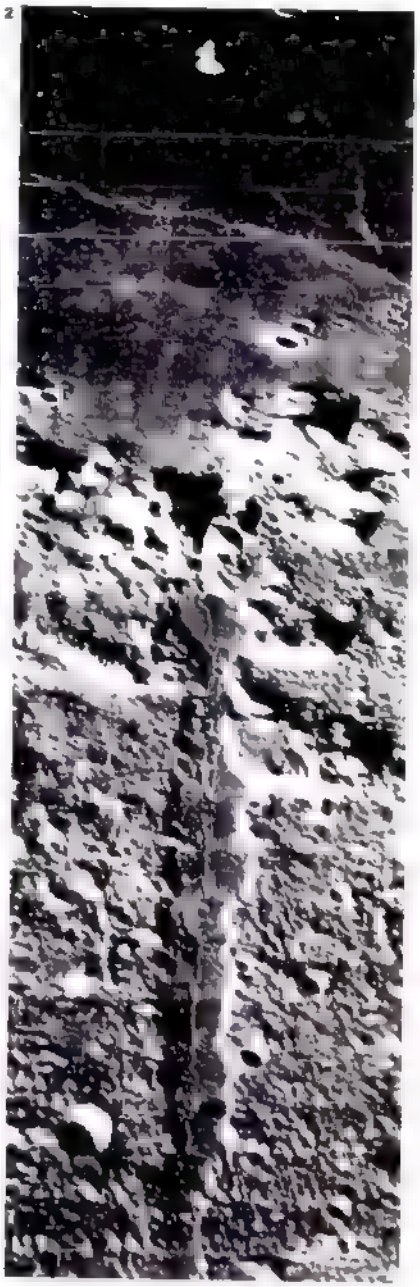
2 The Alpine valley, 130km (80 miles) long, can be seen in this Orbiter photograph. Mont Blanc, the highest peak in the Alps, lies near by. Note the delicate rille running along the valley floor.



3 Aristarchus, with a diameter of 38km (24 miles), is bright compared with the surrounding region. The walls are heavily terraced and the floor contains a central mountain of considerable height.

4 Theophilus, in a photograph taken from Orbiter 3 at an altitude of 55km (34 miles), shows the ramparts and the central mountain mass. Above and to the right can be seen the walls and peak of Cyrillus.





the Orbiter project, involving thousands of photographs – many of which still await complete analysis – there were still a number of outstanding questions. The Apollo programme (which was originally scheduled to continue to number 21) to a large extent supplemented the work of the Orbiters, particularly with more detailed coverage of future landing sites (for example, Apollo 10 the last pre-landing probe, photographed the Mare Tranquillitatis – the site selected for the Apollo 11 touchdown)

Ray-craters

The ray-craters, and the processes by which they may have been produced, also await examination. They must be the youngest of the major features of the Moon, the estimated age of Copernicus and Tycho, for example is no more than one aeon (one thousand million years) but because samples have not been collected and analysed from those regions it is dangerous to be too definite The last Surveyor landed on the outer wall-slopes of Tycho [5] and confirmed that the surface is extremely rough This had been expected, because it was already known from infra-red studies (carried out mainly by the American astronomers J. Saari and R W Shorthill) that Tycho cools down less rapidly during a lunar eclipse, or during a lunar night, than its surrounding areas, this implied a difference in surface covering or texture Other ray-craters behave in the same way and have been called, rather misleadingly hot spots This does not imply any internal source of heat; all it means is that during periods of darkness the "hot spots" record a somewhat higher temperature than other parts of the Moon

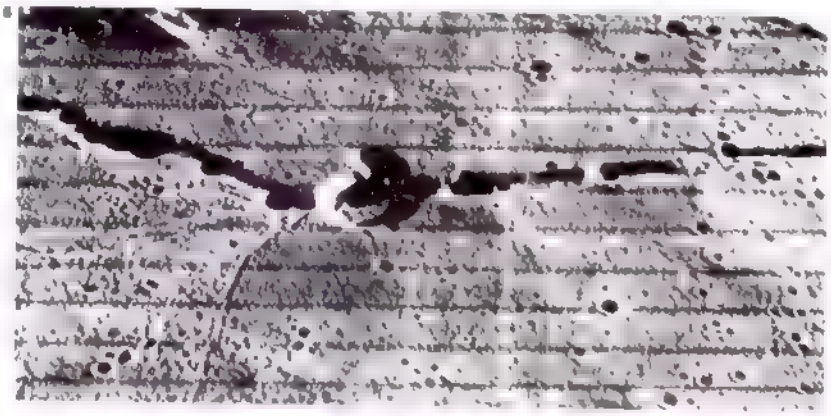
It has been suggested that ray-craters were formed differently from other types of craters, but it seems unlikely. For instance there is no real difference in form between Tycho, which is the centre of the greatest ray-system on the Moon, and Theophilus [4], which is only slightly larger and which is not associated with a comparable system of rays The surrounds are different however; Tycho lies in a crowded area, while Theophilus is a member of a large chain of walled formations

KEY

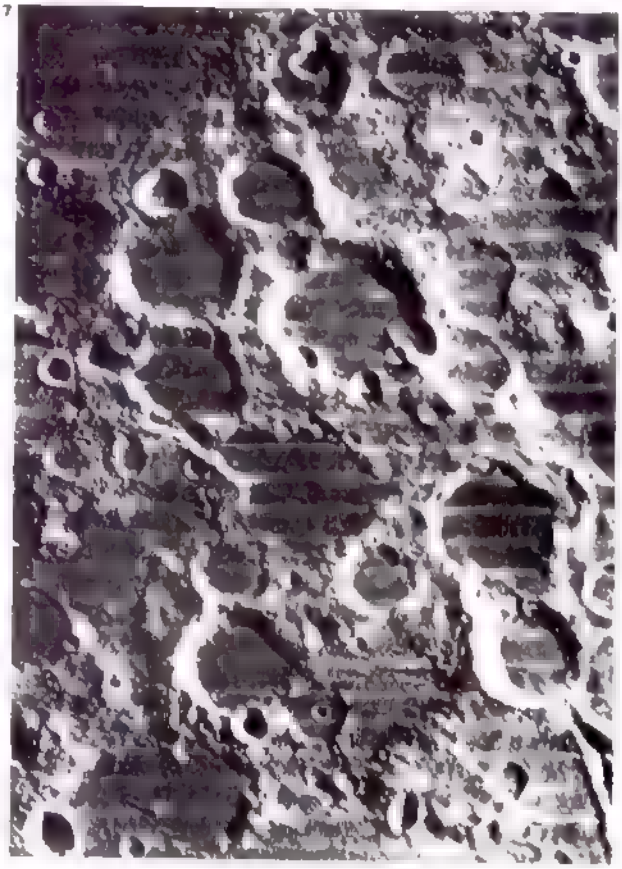
In 1959 the Russian Luna 3 sent back the first close-range photographs of the Moon Mare Moscoviense can be seen

5 Tycho, the great lunar ray-crater, was photographed from Orbiter 5. The terraced walls, the central elevations and the roughness of the floor are clearly shown There is no evidence of lava flows of mare material; Tycho lies in the uplands and is probably one of the youngest of all the major craters

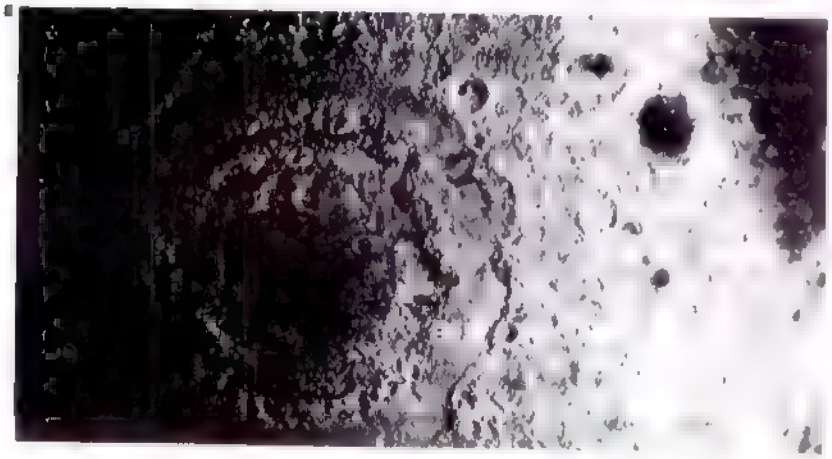
6 The Hyginus Rille is one of the best known on the Moon It is not, properly speaking, a rille or cleft, it is basically a crater-chain as is shown in this view of it taken from Orbiter 3. Hyginus itself, in the mid position in the rille, is 8km (4 miles) across



7 This photograph of the Mare Australe was taken from Orbiter 4 at a height of 3,500km (2,175 miles) above the Moon The craters are flooded with mare material, but Australe is not a regular structure



8 The Mare Orientale, photographed from Orbiter 4, is a complex structure with multiple ring-walls. To the right are seen Grimaldi, which is flooded with more material and Riccioli



The planet Mercury

Mercury can sometimes be seen with the naked eye but is much harder to sight than the other four planets that were known in ancient times – Venus, Mars, Jupiter and Saturn. It is the closest planet to the Sun and takes only 88 Earth-days to travel around it at a mean distance from it of about 58 million kilometres (36 million miles). In size and mass, Mercury is more like the Moon than the Earth [1]. Its diameter is 4,880km (3,032 miles). The escape velocity of only 4.3km (2.7 miles) per second indicates that its atmosphere is negligible.

Observation difficulties

The chief difficulty in observing Mercury is that it is never seen against a completely dark background because it remains in the same part of the sky as the Sun. Although quite bright (at its best exceeding magnitude 0, which is brighter than any star apart from Sirius, Canopus, Arcturus and Alpha Centauri), it is never conspicuous to the naked eye. An observer without a telescope sees it only on favourable occasions, either low down in the west after sunset or low down in the east before sunrise. To make matters worse, the phase (illuminated surface) decreases as Mercury gets nearer to the Earth [3]. At its closest Mercury is at inferior conjunction and cannot be seen at all (except in a rare transit across the Sun) because its dark hemisphere is turned Earthwards.

Mapping the planet

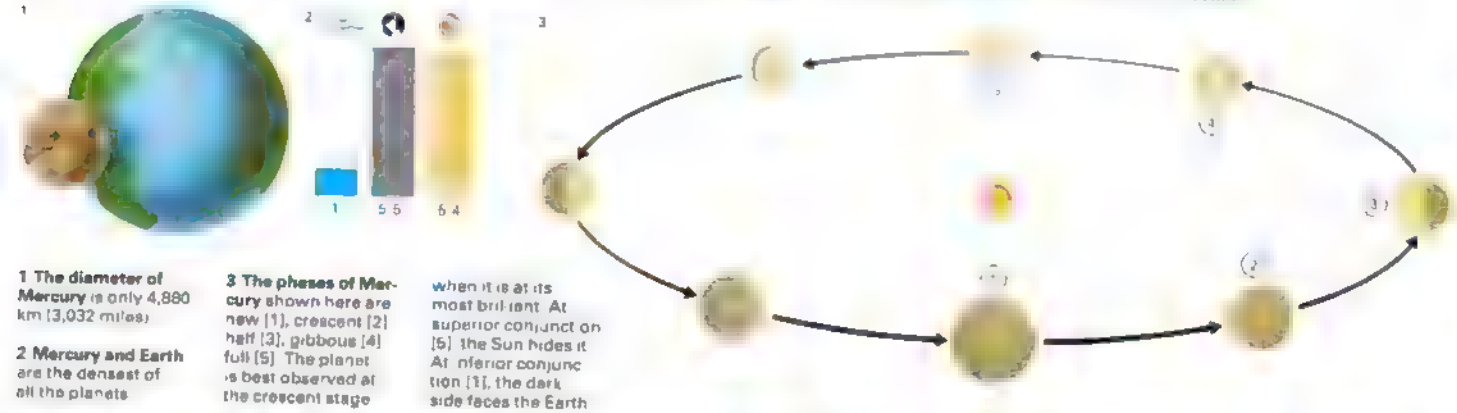
The first serious attempts to map Mercury were made in Milan during the latter part of the nineteenth century by Giovanni Schiaparelli (1835–1910). Rather than study the planet at night when he could see it with the naked eye, Schiaparelli carried out his observations in broad daylight, with Mercury high above the horizon. He was able to see some dark shadings and brighter areas, but his chart was very rough. Later, between 1924 and 1933, a long study was carried out by E. M. Antoniadi. He used the 33in (84cm) refractor at the Observatory of Meudon, making his observations in broad daylight and his chart [Key] remained the best although we now know that it too was highly inaccurate – until the historic flight of Mariner 10 which took place in 1973–4.

Antoniadi, like Schiaparelli, believed that the rotation period of Mercury was captured or synchronous. If that were so, then both the revolution period and the axial spin would be 88 Earth-days and, as a result, part of the planet would be permanently illuminated by the Sun while another part would be in everlasting night. Because the Mercurian orbit is decidedly eccentric, there would be an intermediate "twilight zone" between these two extremes over which the Sun would rise and set, always keeping close to the horizon. The effects would be analogous to the oscillations of the Moon. It is now known that Antoniadi was wrong. Radar measurements carried out since 1962 show that the true rotation period is 58.7 Earth-days so that all regions of the planet receive sunlight at some time or other [5]. Because of a curious relationship (which may or may not be coincidental), Mercury presents the same hemisphere to the Earth every time it is best placed for study, this is what misled the earlier observers.

Antoniadi believed that he had observed local obscurations or "clouds" on Mercury.

CONNECTIONS

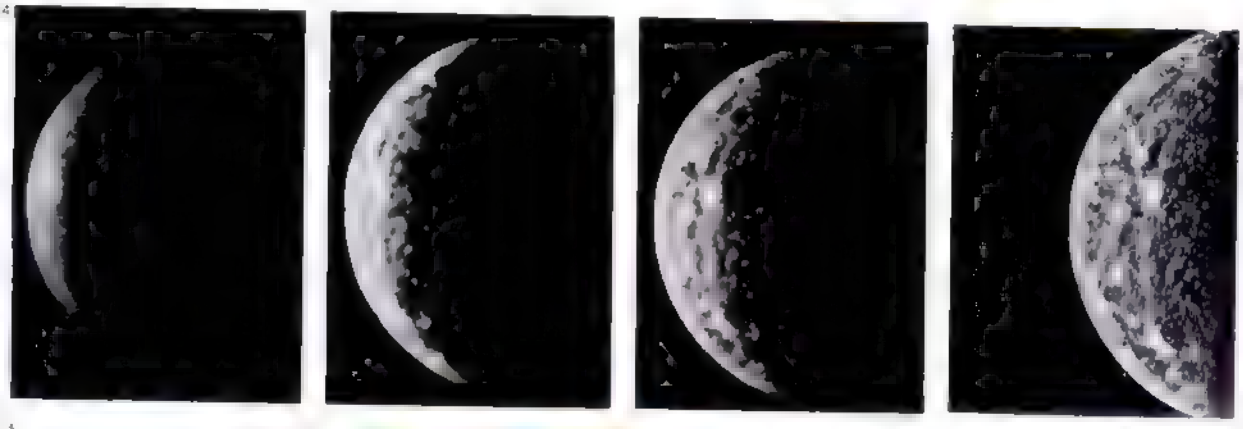
See also



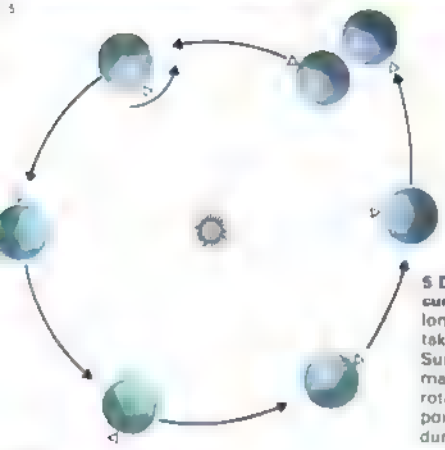
1 The diameter of Mercury is only 4,880 km (3,032 miles)

2 Mercury and Earth are the densest of all the planets

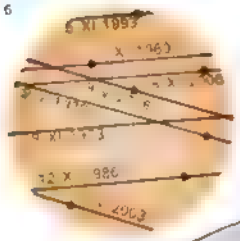
3 The phases of Mercury shown here are new [1], crescent [2], half [3], gibbous [4], full [5]. The planet is best observed at the crescent stage when it is at its most brilliant. At superior conjunction [5] the Sun hides it. At inferior conjunction [1], the dark side faces the Earth



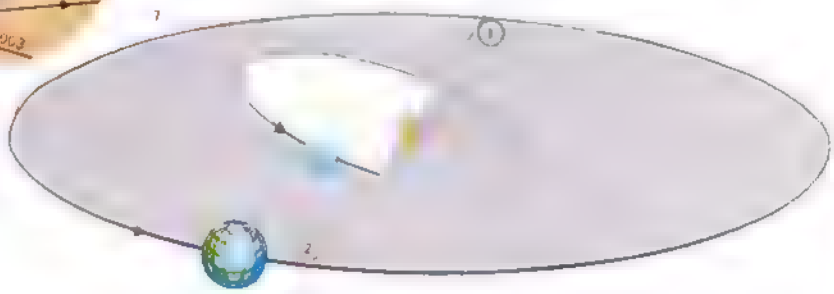
4 Four photographs taken from Mariner 10, in March 1974 show Mercury in different phases. Surface features became clearer as Mariner approached. The resolution on the last picture is 20km (12.4 miles) – the size of the smallest discernible object. The probe was then 952,600km (591,900 miles) from the planet. Craters show up best when close to the Mercurian terminator (boundary between the illuminated and dark parts)



5 Daylight on Mercury lasts about as long as the planet takes to orbit the Sun – 88 days. The marker shows the rotation of a fixed point on the planet during this period



6 Transits of Mercury (its passages across the Sun as seen from Earth) are plotted from 1960 to 2016. Not all are equally long, that of 2016 will last much longer than that of 2003. During a transit Mercury shows a small, well defined black disc, but is invisible to the naked eye



7 Transits are rare as Mercury's orbital angle is different from the Earth's. They usually occur in November [1] or, less often, in May [2]

which sometimes veiled the surface features. This was puzzling because it was obvious that the atmosphere must be extremely rarefied by terrestrial standards. It is now known that the atmosphere is much too thin to support clouds of any kind and there seems no escape from the conclusion that Antoniadi's observations were wrong. Because of Mercury's closeness to the Sun, its surface is extremely hot during daylight and a thermometer would register a temperature of more than 370°C (700°F) at maximum; but because of the virtual lack of atmosphere the nights are bitterly cold. No known form of life could survive on Mercury.

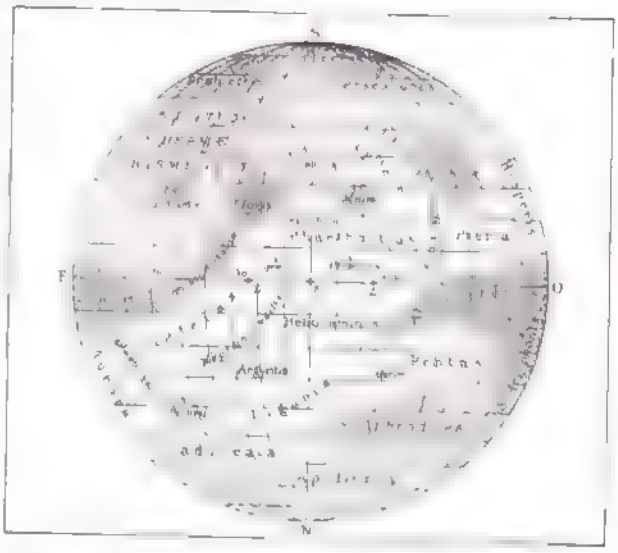
Mariner 10 probe

The first close-range information about Mercury's surface was received in 1974 from the flight of Mariner 10, the first "two-planet" probe. In February 1974 it bypassed Venus, sending back pictures, and then swung towards towards a rendezvous with Mercury during the following month. The photographs revealed a landscape that was strikingly like that of the Moon [4]. Craters, mountains, ridges and valleys showed up everywhere, although there were fewer broad dark plains like the lunar Mare Imbrium. On Mercury, the chief plain has been named the Caloris Basin [9].

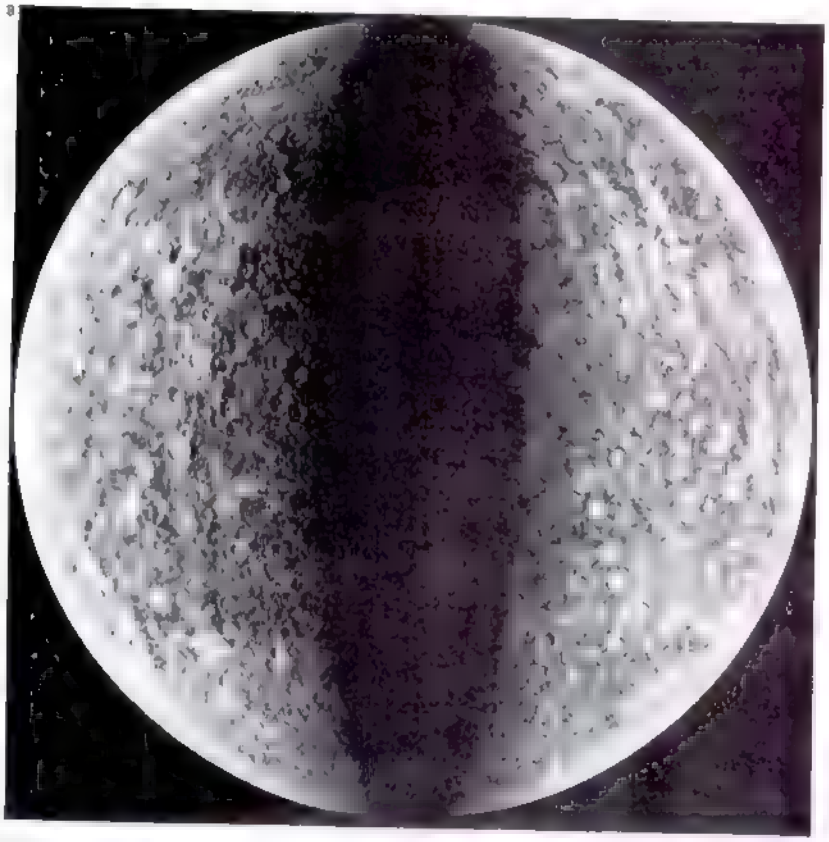
In September 1974, after orbiting the Sun, Mariner 10 made a second rendezvous with Mercury and took more high-quality pictures. The third encounter took place in March 1975. Despite the incompleteness of the pictures, they revealed the same types of craters and mountains. Much of Mercury has now been mapped and for the first time there is reliable information on what this strange world is really like.

One discovery of great interest was the detection of a magnetic field. This field is weak compared with that of the Earth, but it is quite definite and produces a true magnetosphere.

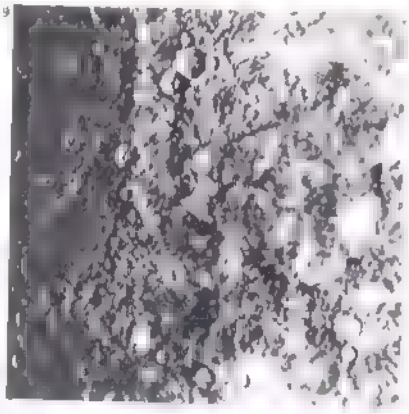
Further probes to Mercury are being planned, but the chances of manned expeditions there in sight, at least in the foreseeable future. Mercury is overwhelmingly hostile, although from the scientific point of view it is of exceptional interest.



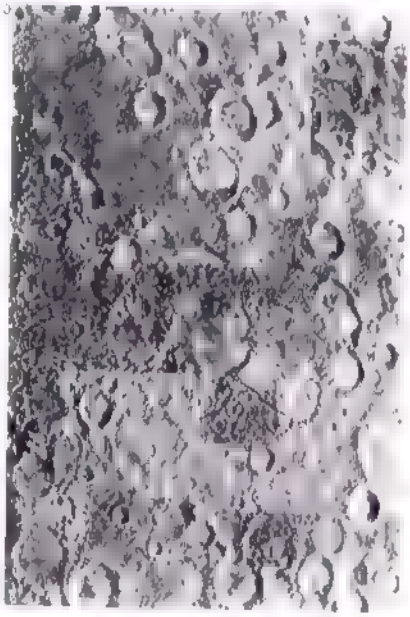
This map of Mercury was drawn between 1924 and 1933 by Antoniadi, who named the main features. There is poor correlation between the dark areas shown and the results from Mariner 10, and some recasting of names will be needed.



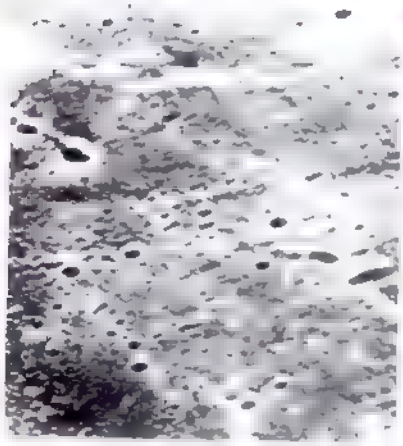
8 The whole surface of Mercury shown in this mosaic of Mariner 10 photographs is extensively cratered and distinctly lunar in appearance. There are also bright ray systems (upper right), like those of the Moon. The Mariner photographs were highly informative.

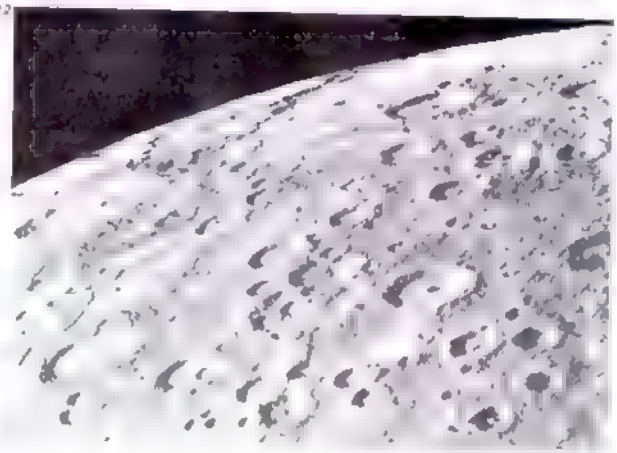


9 This mosaic of the Caloris Basin from Mariner 10 shows what is probably the most distinctive plain on Mercury. The mountain rim around the basin can be clearly seen and the interior is different from the surrounding region. The origin of the basin is still unknown.



10 Hills and ridges cut across many of the craters in the region shown here, which is unique to the planet. Their origin is obscure. They may be the result of shock waves caused by the formation of the 1,300km (800 mile) Caloris Basin, which lies diametrically opposite this region.





11 Another region of Mercury shows an area crowded with craters and craterlets. One interesting feature is the crater valley to the upper left, which resembles many of the crater chains on the surface of the Moon.

12 The similarity of Mercury to the Moon is clear in this Mariner 10 photograph. Some craters have centred peaks and all the well marked craters are basically circular, although here they appear elliptical because of foreshortening.

The planet Venus

Venus, the second planet in order of distance from the Sun, is almost as large as the Earth and has more than 80 per cent of the terrestrial mass. [1, 2]. Instead of being devoid of atmosphere, it has a deep, dense and cloudy surround which prevents us from ever seeing its true surface.

The mean distance of Venus from the Sun is 108 200,000km (67,200 000 miles) and thus is practically constant because its orbit is more closely circular than that of any other planet. The revolution period is 224 7 days. Before the age of space probes and powerful radar the rotation period of Venus was unknown, there are no markings on the disc persistent enough to be used for rotation measurements.

Observing Venus from Earth

To the naked eye Venus is a splendid object and is far brighter than any other celestial body apart from the Sun and the Moon, which is why it was named after the goddess of beauty. Telescopically, however, Venus is a disappointment and it is not therefore surprising that until recent years it was often called the "planet of mystery" [4]

In the 1930s some positive information emerged. It was established that the atmosphere of Venus is made up largely of carbon dioxide, which tends to act as a "blanket" and shut in the Sun's heat. Later Earth based observations showed that the clouds are made of sulphuric acid.

In 1962 astronomers bounced radar pulses off Venus, in the first interplanetary radar experiment. Radar pulses can penetrate cloud, so the returning echo was received from the unseen surface of Venus below the clouds. The first discovery was a surprise: the axial rotation period is slow - about 243 Earth days, which is longer than the revolution period of 224.7 days, therefore on Venus the day is longer than the "year" giving rise to a peculiar calendar.

It has now been established that Venus rotates east to west, instead of west to east as with the Earth.

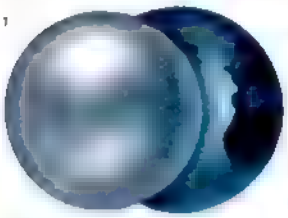
Information from early probes

In 1962, the US probe Mariner 2 passed Venus and discovered that its surface is extremely hot. This was confirmed by later Soviet landers, which also found that the carbon dioxide atmosphere produces a very high surface pressure. On its way to Mercury another US probe, Mariner 10, photographed the cloud tops of Venus in 1974, and confirmed earlier suspicions that they rotate much faster than the planet itself, with a period of only four days.

The next step came in October 1975 when two Soviet probes, Veneras 9 and 10 made controlled landings on the surface and transmitted pictures back. The pictures were relayed by the orbiting sections of the probes, which remained circling Venus at a height of about 1,500km (900 miles). It was an amazing triumph for the Soviet scientists - even though neither Venera 9 nor 10 could transmit for more than about one hour before being put permanently out of action by the extreme temperature and pressure.

Surprisingly, the surface of Venus proved to be strewn with smooth rocks, many of which were about 1m (39in) in diameter [12]. There was plenty of light – in the Russian description, about as much as at

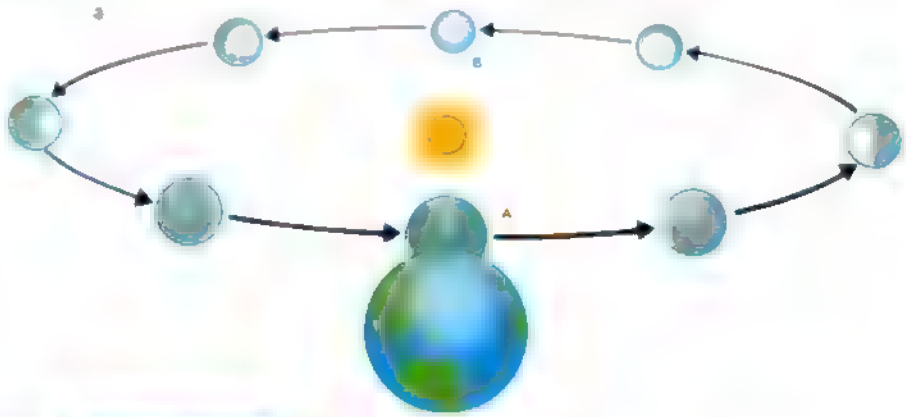
CONNECTIONS

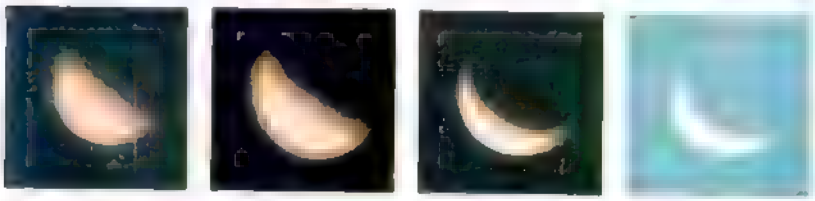


1 Venus and the Earth are almost equal in size, mass and surface gravity

2 The density of Venus is less than Earth's, but it may have a heavy core

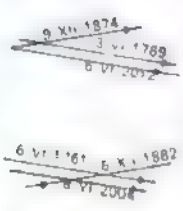
3 Venus is brightest when at the crescent stage. When the planet is new [A] it is invisible (except at transit). At superior conjunction and full phase [b] it is on the far side of the Sun





4 The apparent diameter of Venus changes according to phase: it is least at superior conjunction, when Venus is full, because the planet is then on the far side of the Sun and is at its most remote from the Earth. The apparent diameter increases as the phase shrinks, as is shown in this superbly clear series of photographs which were taken with a (12in) 30cm reflecting telescope

5 Mariner 10 provided the first really conclusive optical evidence of the four-day rotation of the upper clouds of Venus (although this phenomenon had been suspected from observations made from Earth) with these three photographs taken in February 1974. The cloud patterns are clearly shown (the arrow indicates the same area of Venus). The first photograph was taken on 2 February at 0 hours, the second at 7 hours and the third at 14 hours (note the movement of the arrowed marking). The axis of Venus is almost perpendicular to the plane of the orbit, but the rotation is retrograde in direction



6 Transits of Venus Its passage across the Sun as seen from Earth occur in pairs with an interval of eight years, each pair separated by more than a century. The diagram shows the transit paths of Venus for the years 1874-82 and 2004-12

7 When Venus enters transit, it seems to draw a strip of darkness (Black Drop) after it, an effect produced by the planet's atmosphere. The strip disappears only when the planet is well advanced on to the Sun's disc

8 The Ashen Light is a faintly luminous patch that appears on the night side of Venus when it is a crescent. This is not the same as Earthshine on our Moon, and its cause is uncertain. Some authorities regard it as a contrast effect; others believe it to be due to electrical phenomena in the upper atmosphere of the planet. It is not easy to observe, but most serious students of Venus have seen it. It was first recorded in the 1790s. (This drawing is exaggerated)

noon in Moscow on a cloudy summer day. A temperature of 485 C (900 F) was recorded and pressure was found to be 90 times that on Earth.

Pioneer unveils Venus

In December 1978, several US and Soviet probes reached Venus, some to land and some to go into orbit. These probes – particularly the US Pioneer Venus Orbiter – have revolutionized our knowledge of Venus.

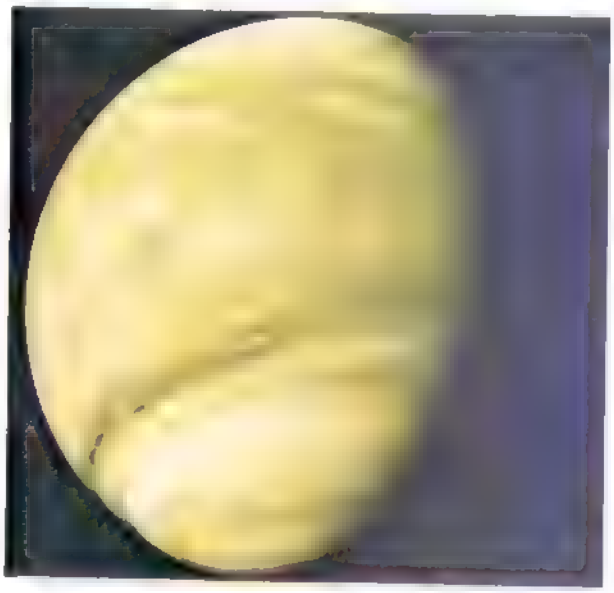
The Pioneer Orbiter carried a small radar set, with which it could map the surface in detail. Most of Venus consists of rolling plains, but there are two high plateaux. Aphrodite Terra, near the equator, is about half the size of Africa. Ishtar Terra, in the northern hemisphere, is rather smaller, but includes the highest peak on Venus, Maxwell Montes, which is 10,800m (35,400ft) above the average level of Venus – higher than Mount Everest. Venus also has a pair of adjacent volcanoes probably extinct, Rhea and Theia, which rise to 4,000m (13,000ft).

Because Venus is similar in size and composition to the Earth, many geologists had expected it to show signs of the "continental drift" which moves blocks of the Earth's crust about. Pioneer Venus has found some rift valleys in the surface of the planet – one of them 280km (170 miles) wide and 2,250km (1,400 miles) long – but no sign of the great ridges and mountain chains which result from continental drift on Earth. This is a major puzzle.

The high surface temperature is maintained by sunlight and heat trapped in the atmosphere: carbon dioxide traps more than half the total amount, water vapour about a quarter, and clouds, haze and sulphur dioxide account for the rest. Despite its important heat-blanketing effect, water vapour is extremely scarce in the atmosphere of Venus.

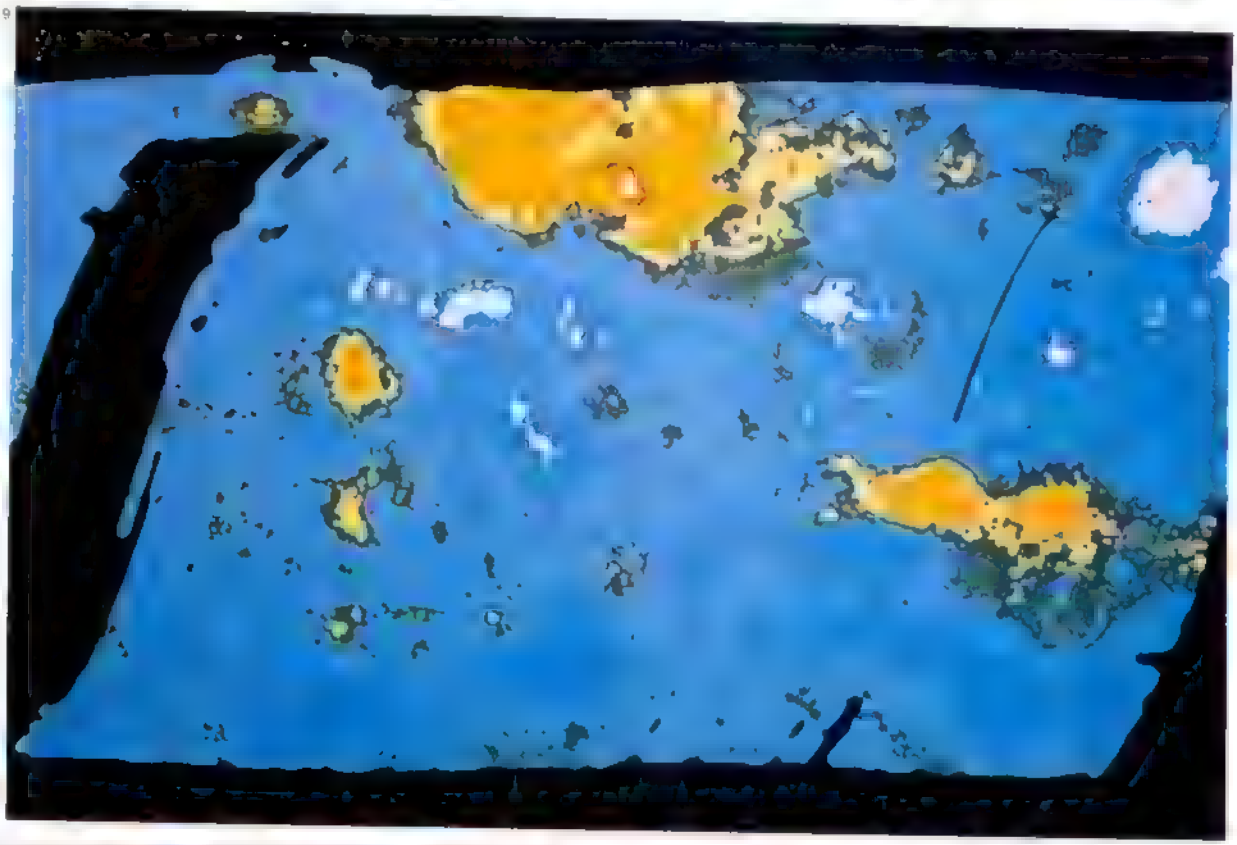
Venus's main cloud layers lie at a height of 50km (30 miles), with patchier smog layers above and below. There are high speed winds within and above the cloud layers, reaching 360 km/h (225 mph) at a height of 65km (40 miles). Lightning is common, especially above its volcanoes, and the upper atmosphere glows at night because of chemical reactions in its gases.

KEY



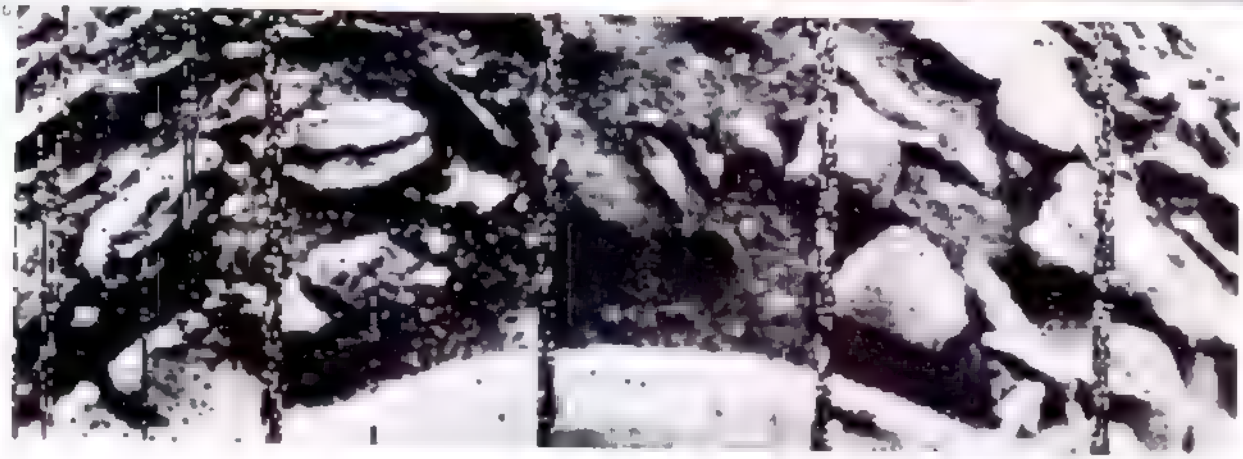
The US Pioneer Venus probe went into orbit round Venus in 1978 and sent back detailed views of its clouds.

9



9 A contour map of Venus, obtained by radar, was one of the outstanding achievements of the US space probe Pioneer Venus Orbiter. This Mercator projection covers about 80 per cent of the planet. The contours are at intervals of 500m (1,640ft), and the black stripes are regions still to be mapped. Sixty per cent of Venus is covered by rolling plains (dark blue), and 16 per cent by lower basins (pale blue). Venus has three main highland regions (orange). Beta Regio (left) consists of two adjacent 4,000m (13,000ft) volcanoes, Rhea and Theia. Aphrodite Terra (right) is a plateau half the size of Africa. Ishtar Terra (top) is smaller, about the size of Australia (although it seems larger in this projection). It contains Maxwell Montes (white), the highest peak on Venus, at 10,800m (35,400ft).

10



10 The first view of the surface of Venus was sent back by the Soviet probe Venera 9 on 22 October 1975. It shows rough angular rocks stretching to the horizon (crossing the top corners at an angle). Part of Venera can be seen at the bottom. Its detached density probe is arrowed. The scene was surprisingly bright, "like a cloudy winter day in Moscow". Venera 10 landed three days later and found a wider panoramic rocks.

The planet Earth

The Earth is the largest member of the group of inner planets and is also the most massive. The difference in size and mass between the Earth and Venus is slight (the ratio for mass is 1 to 0.82), but Mars is much smaller, and Mercury is more comparable with the Moon than with the Earth.

When the Earth is compared with its planetary neighbours, marked similarities as well as marked differences are found. Of course, what singles the Earth out from any other planet is the fact that it has an oxygen-rich atmosphere and a temperature that makes it suitable for life of the kind we can understand. Were the Earth slightly closer to the Sun, or slightly farther away, life here could not have developed.

The ecosphere

What is termed the "ecosphere" [4], or the region in which solar radiation will produce tolerable conditions for terrestrial-type life, extends from just inside the orbit of Venus out to that of Mars. Until about 1960 it was commonly thought that such life might exist throughout the whole region. Although this possibility was far more remote for Mars with a significantly lower mass than that of Earth, and hence a tenuous atmosphere, Venus was looked upon as the Earth's twin. Approximately equal in density as well as size and mass, Venus also absorbs about the same amount of solar energy as Earth because of the high reflecting power of its cloud. It was not until 1967 when the surface temperature of Venus was shown to register up to 485°C (900°F) that it was commonly accepted that advanced terrestrial life could develop only within a limited zone.

Another essential need is the presence of an atmosphere that will not only enable living creatures to breathe, but will also protect the planet from lethal short-wave radiations coming from space. On the surface of the Earth there is no danger, because the radiations are blocked out by layers in the upper atmosphere, but the Moon is unprotected, and so too is Mercury. Had the Earth been more massive, it might have been able to retain at least some of its original hydrogen (as Jupiter and Saturn have done) and the resulting atmosphere might have been unsuitable for life. A lower mass might have led to the escape not only of the hydrogen but also of all other gases, so that a fortunate combination of circumstances has produced terrestrial life.

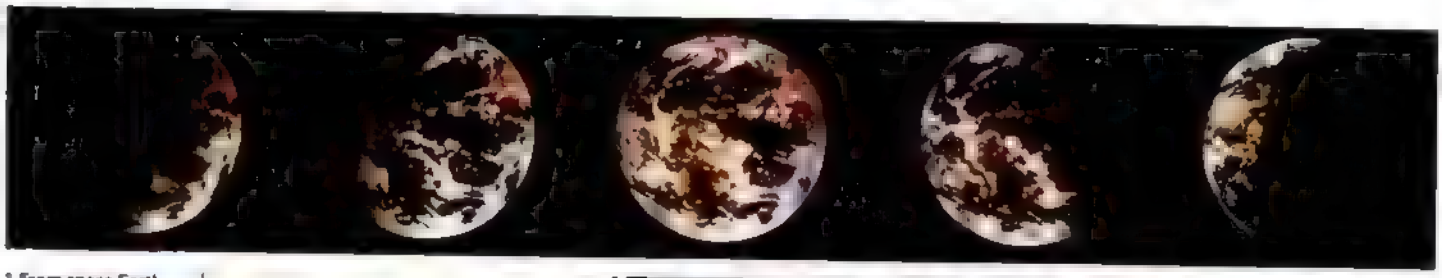
Next, there is the question of temperature, which does not depend only upon the distance of the planet from the Sun or upon the composition of its atmosphere; there is also the axial rotation period to be taken into account. The Earth spins round once in approximately 24 hours, and the rotation period of Mars is only 37 minutes longer, but the situation with Mercury and Venus is very different: the periods are 58.7 days and 243 days respectively, leading to very peculiar calendars. Were the Earth a slow spinner, the climatic conditions would, as a result, be both unfamiliar and hostile.

The Earth's magnetic field

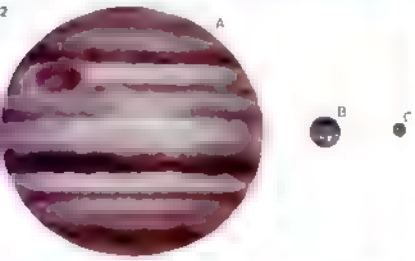
The Earth's heavy, iron-rich core is associated with the magnetic field, and here too comparisons may be made with other planets. Venus again provides problems. With its comparable size and mass it should

CONNECTIONS

See also



1 From space Earth will show phases, just as the Moon does to us. The five photographs shown (A) were taken from a satellite over a period of 12 hours. B shows the phases of the Earth as they are seen from the Moon. For this purpose it can be assumed that the Earth is stationary, with the Moon moving round it in a period of 27.3 days. When the Moon is full from Earth (1) a lunar observer will have "new earth" at our new Moon (3) a lunar observer will see full Earth. Below the main diagram, the different phases of the Moon (left) and Earth are shown (1–4).



2 Relative sizes of Jupiter (A), the Earth (B) and Mercury (C). Jupiter is the largest planet, Mercury the smallest, Earth is intermediate in size, but much more nearly comparable with Mercury. Earth is in fact the largest of the so-called terrestrial planets (Mercury, Venus, Earth, Mars, Pluto) but far inferior to the smallest of the giants (Uranus).

3 Apollo 10 sent back these pictures in May 1969. It was the second vehicle to take men round the Moon. The pictures show the Earth coming into view as the spacecraft comes from behind the far side of the Moon, from which the Earth can never be seen. In the first photograph the sharpness of the lunar horizon is particularly notable, there is no lunar atmosphere to cause the slightest blurring or distortion. Note the changing Earth phase.

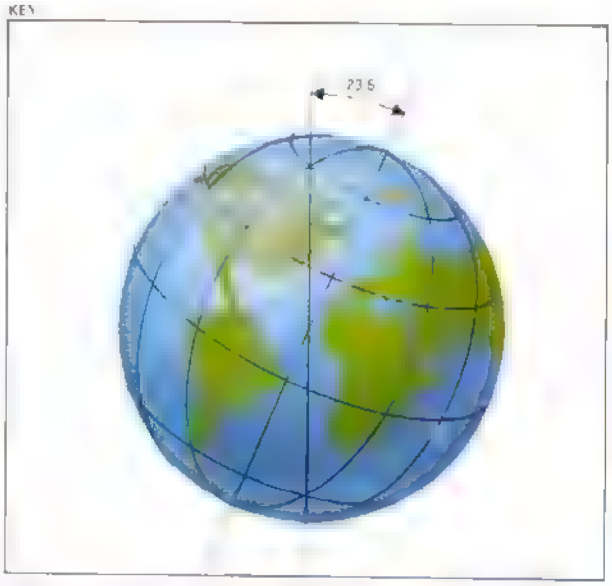
have the same kind of core and hence an appreciable magnetic field; but space probes have so far failed to detect any magnetism, and it is now certain that even if a magnetic field exists it is very weak indeed. The same may be true of Mars, but Mercury has a perceptible field and even a magnetosphere. It is probably significant that Mercury and the Earth are the densest of all the planets, with specific gravities of about 5.5 (that is to say, their masses are 5.5 times greater than that of equal volumes of water).

The watery planet

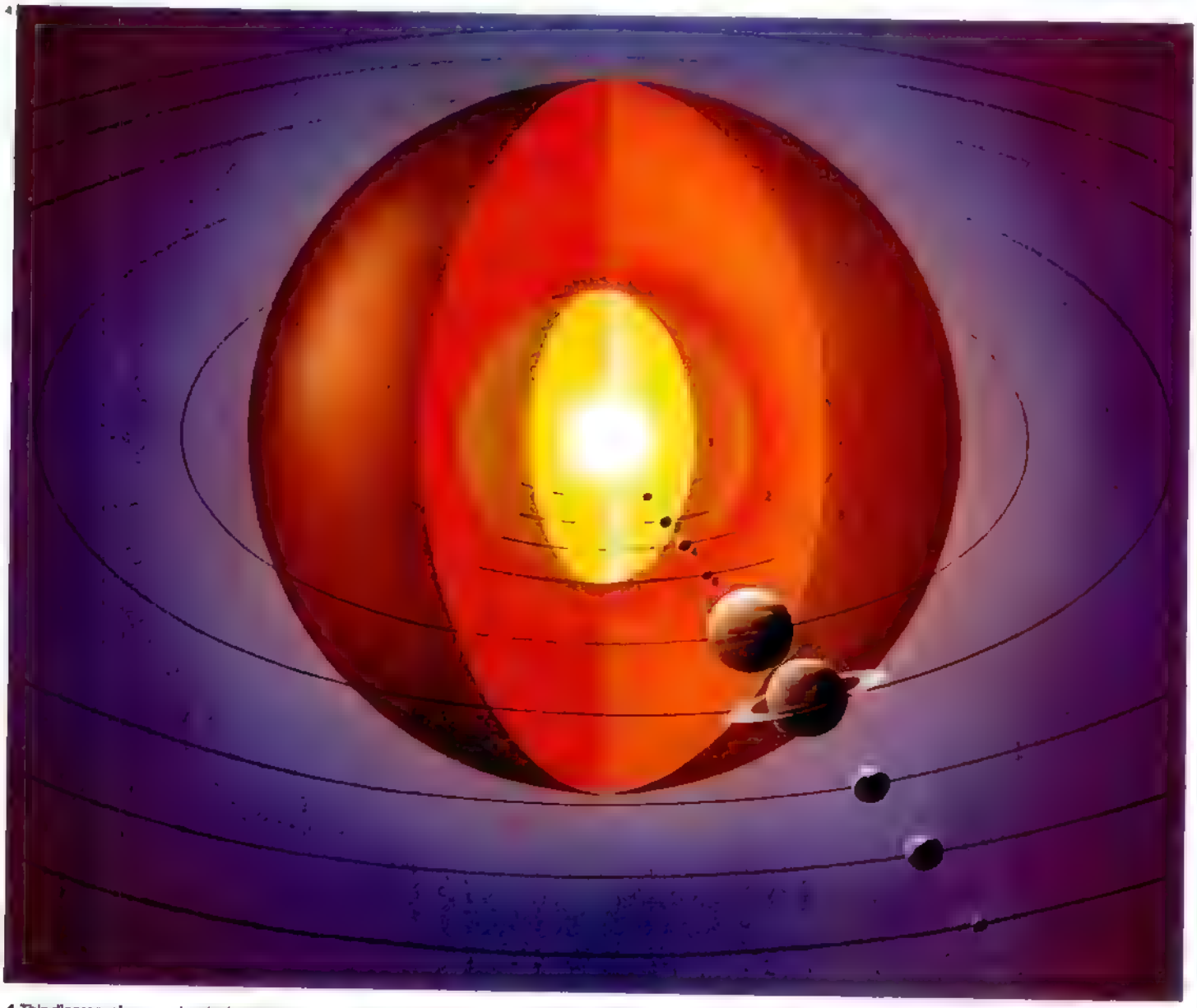
The Earth is again unique in having a surface that is largely covered with water; thus although it is the largest of the four inner planets its land surface is much less than that of Venus and equal to that of Mars. There can be no oceans or even lakes on Mars, because of the low atmospheric pressure, and of course none on the Moon or Mercury, which are to all intents and purposes without atmosphere. On Venus the surface temperature is certainly too high for liquid water to exist, so that the old, intriguing picture of a "Carboniferous" type Venus, with luxuriant vegetation flourishing in a swampy and moist environment, has had to be given up.

Because the Earth is so exceptional, there have been occasional suggestions that it was formed in a manner different from that of the other planets, but this is certainly not so. The age of the Earth, as measured by radioactive methods, is approximately 4,600 million years (4.6 aeons) and studies of the lunar rocks show that the age of the Moon is the same; there is no reason to doubt that the Earth and all other members of the Solar System originated by the same process, and at about the same time, from the primeval solar nebula. It is often said that Mars is more advanced in its evolution than is the Earth and this may be true; but the absolute ages of the two worlds are probably the same, although Mars has "aged" more quickly.

The Earth's position in the middle of the ecosphere, and the particular size and mass that lead to its own kind of atmosphere, single it out. There is no other planet in the Solar System upon which men could survive except under artificial conditions.



The Earth's axis is inclined at 23.5° to the perpendicular to the orbital plane. This causes the seasons, the varying distance of the Earth from the Sun has only a minor effect.



4 This diagram shows the ecosphere, or the region round the Sun in which a planet will be at a suitable temperature for life to exist – assuming that the planet is of Earth type. The inner yellow zone (1) is too hot. Beyond is the ecosphere (2) (orange), and beyond this (3) is the zone where temperatures will be too low. The Earth (4) lies in the middle of the ecosphere, Venus (5) is at the inner limit and Mars (6) at the outer

The planet Mars

Mars, the first planet beyond the Earth in the Solar System, is of especial interest to man. In the earlier part of this century many astronomers believed implicitly in the existence of an advanced civilization on Mars. This belief has now been shattered. There are no Martians and it seems that the most advanced life the planet could support would be very primitive organic matter. It is more likely that the planet is sterile. Despite this, Mars is still more Earth-like than any other known world and it must surely be the first target for an expedition by a manned space probe after the Moon.

Through a telescope, Mars shows a red disc with white caps at the poles and prominent dark markings that are essentially permanent [Key]. Its mean distance from the Sun is 228 million km (141.5 million miles). The Martian year is equal to 687 Earth-days and a day is equal to 24hr 37min. Moreover the axial tilt is only slightly greater than the Earth's, so that the seasons are of the same basic type, although much longer. As with the Earth, the south pole is tilted Sunwards at perihelion (the point in orbit when a planet is nearest to the Sun); the effect of this on the climate of Mars is greater than its effect on Earth's climate because the orbit of Mars is more eccentric. The climates in the planet's southern hemisphere are more extreme than those in the northern; the summers are shorter and hotter and the winters longer and colder. The noon temperature on the Martian equator at midsummer may rise to more than 16°C (60°F). The nights are bitterly cold, because the thin atmosphere is inefficient in retaining warmth, however, Mars is certainly not a frozen world.

The Martian atmosphere

Because Mars is not only less dense than the Earth but much smaller – the diameter is a mere 6,790km (4,220 miles) – the escape velocity (the speed an object must reach to overcome gravity) is low: 5.0km (3.1 miles) per second, which explains the tenuous atmosphere. The main constituent is now known to be carbon dioxide (95 per cent), and the barometric pressure at ground level is less than 10 millibars. No advanced terrestrial creatures could survive without protection.

There is no free water on the Martian surface today. It appears, however, that the white polar caps are made up chiefly of water ice with some carbon dioxide (dry ice) also present. The size of the caps varies according to the Martian season, at their greatest extent they are conspicuous and can be seen with a small telescope [7].

Features shown by early maps

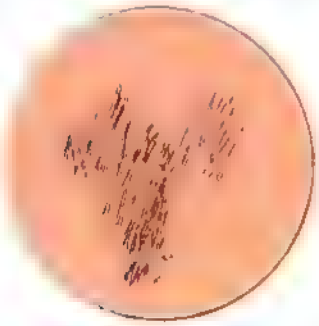
The first drawing to show any markings on the surface of Mars was made by the Dutch astronomer Christiaan Huygens (1629–95) in 1659 [1A]. It shows the V-shaped region now called the Syrtis Major Planitia, although in somewhat exaggerated form. Later observers using more powerful telescopes produced drawings that showed more detail. The first reasonably reliable maps date from the second half of the nineteenth century. What may be called the "modern" period of telescopic research was initiated by Giovanni Schiaparelli (1835–1910) in 1877 when Mars was at perihelion and opposition and was excellently placed for observation.

Schiaparelli, observing from Milan, drew famous canal network appears in Schiaparelli's drawing [C] made in 1877. E. M. Antoniadi, who used the 33in (83cm) Meudon refractor from 1900–30, had no faith in the canals, but his drawing [D] was remarkably accurate. The drawing by Percival Lowell, made in about 1905 [E], shows the illusory canal network.

CONNECTIONS

See also

1A



B

C

D

E

1 Different stages in the telescopic exploration of Mars are shown in these five drawings. The first [A], made in 1659 by Christiaan Huygens, shows only the Syrtis Major. In drawing B, made by the pioneer observer Johann Schröter (1746–1816) in 1800, the Syrtis Major is again shown reasonably accurately. The

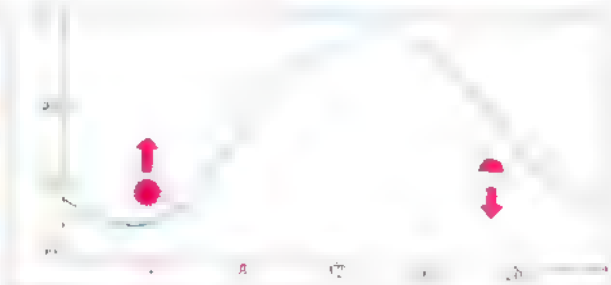
2

2 The surface area of Mars is 28% that of the Earth. Its diameter of about 6,790km (4,220 miles) is a little more than half that of the Earth and approximately twice that of the Moon. It has only one tenth of the Earth's mass.

3

3 The density of Mars is appreciably less than the Earth's with a specific gravity of only 3.94, resulting in a fairly low escape velocity. The Martian surface gravity is 0.38 that of the Earth's. There is no detectable magnetic field and so presumably no heavy core.

4



B

4 Temperature and wind speed on the Martian surface were carefully monitored by Viking 1 in the area of Chryse Planitia. Despite a diurnal range of more than 70°C [A], the maximum temperature recorded during a Martian day was still below freezing-point. Chryse is, however, well to the north of the Martian equator. Speed (metres per second) and direction of winds were also recorded during the Martian day [B]; wind direction is indicated by the arrows. The pattern was repeated day after day, the winds remaining gentle, never attaining 24km/h (15mph) throughout the period. Viking's meteorology station was set up, however, when conditions in its subtropical location were very steady.

a map that was superior to any of its predecessors [1C]. On it he showed straight, artificial-looking features that he called *canali* or channels but which have since become known as the Martian canals. Inevitably, it was suggested that these features were artificial waterways constructed by the planet's inhabitants to form a vast irrigation system. According to this intriguing theory, water would be drawn from the ice-covered polar caps and pumped through to the arid regions closer to the equator. Where canals crossed each other, they did so with the formation of small patches called "oases" which were regarded as centres of population. Schiaparelli himself kept a reasonably open mind. But the American astronomer Percival Lowell (1855–1916), the founder of the Lowell Observatory at Flagstaff, Arizona, USA, was convinced that Mars supported a highly developed civilization.

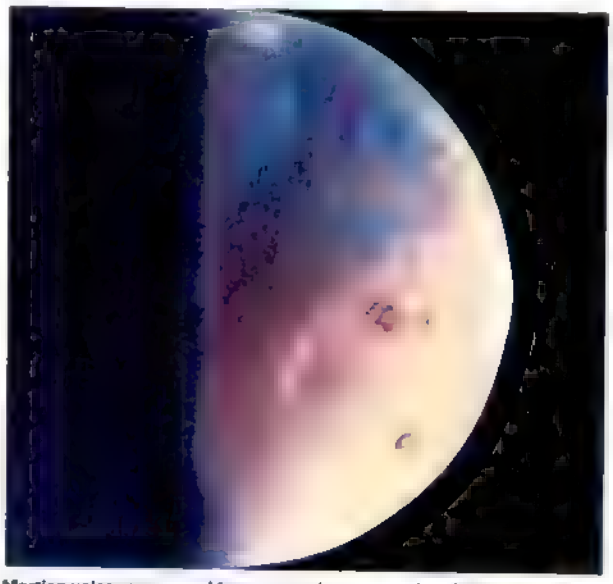
The Martian surface

After it became obvious that dark patches on the surface of Mars could not be seas, it was generally believed that they must be tracts of vegetation growing in depressions. This was still thought to be true until Mariner 4, the first successful space probe to Mars, was launched in 1965. It has now been found that the dark regions are not depressions, some of them, including the Syrtis Major, are lofty plateaus sloping off to all sides. Evidently their colouring arises from some difference in surface texture and not from any vegetation.

Most of the Martian surface is reddish-ochre; these areas are generally called deserts, which may be an appropriate name although there is no real analogy with terrestrial deserts such as the Sahara. Dust-storms are not uncommon and there are winds in the atmosphere of the planet [6].

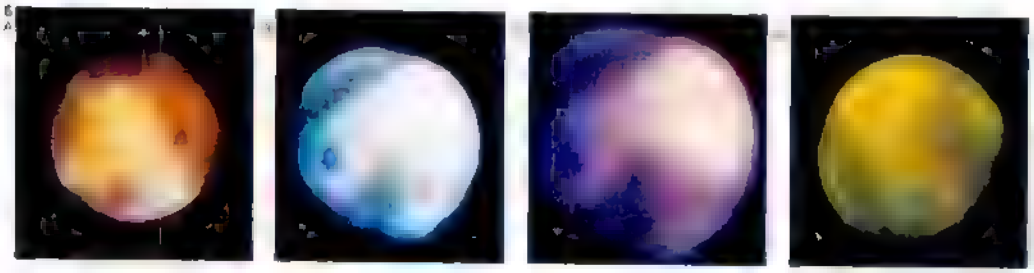
While the Martian "canals" do not exist, Viking spacecraft have given evidence that water may once have flowed on Mars in abundance. In the Chryse region cleanly cut channels meander and intertwine like dried-out river beds, and "islands" of ancient rock have tails extending downstream. According to a later theory, the mysterious channels may have been carved by molten lava flowing over the rock.

KEY



Martian volcanoes, three in the Tharsis Mts, plus Olympus Mons, were photographed by Viking 1 from 560,000 km (348,000 miles). Also seen is the impact basin Argyre.

6 A



5 Martian clouds consist of high-altitude "white" clouds and much less common but more extensive dust clouds. These four views show Mars taken through different coloured filters. The area of cloud that forms every summer over the Syrtis Major is shown in A, B, C and D. In D the Syrtis Major to the left, covered by the same cloud. The bright area to the right is Elysium.

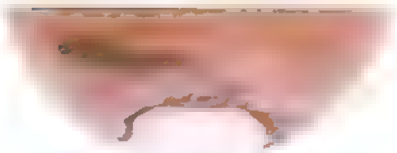
6 The Martian dust-storm of 1971 was one of the greatest ever observed. [A] 20 Sept 1971, before the storm, the dark markings show up clearly. [B] 12 Oct 1971, the dust covers the planet. Mariner 9 approached Mars during this period. [C] 8 Feb 1972, the dust is clearing and the most prominent surface features become visible.

7 Observations of Mars made in 1972 show a dark peripheral band well defined in A at the edge of shrinking polar caps. B to F, Gerard de Vaucouleurs in 1939 attributed the shrinking to release of water vapour, which might support vegetation. Vegetation is now considered unlikely although the polar caps do contain water ice.

7 A

22 May, 02 10 hrs

B



21 June, 01 30 hrs

C

6 July, 00 20 hrs

D

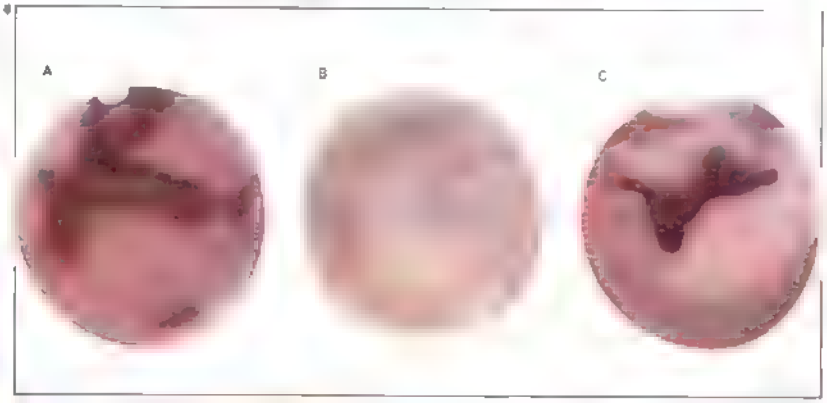
16 August, 22 00 hrs

E

28 Sept, 20 15 hrs

F

10 Oct, 19 45 hrs



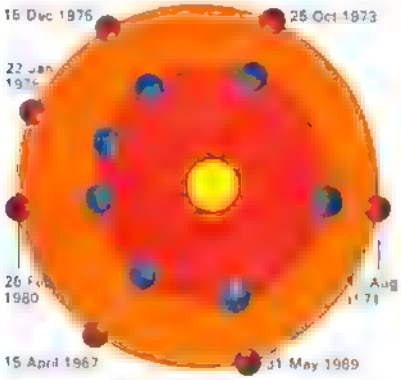
B A

B

C

D

E



8 The "opposition of Mars" describes the lining up in sequence of the Sun, Earth and Mars. Mars takes 687 Earth days to complete one revolution round the Sun and this means that it comes to opposition once in approximately 780 days. As shown in E, oppositions occurred in 1967, 1969, 1971, 1973 and 1975; the next will be in Jan 1978 and Feb 1980. The series of drawings [A to D] begin with Earth and Mars in opposition. By the time Mars has reached position 1 [A], the Earth will have moved to position 2. By the time the Earth has completed one full circuit [B], Mars has made only a little more than half a revolution. When the Earth has made 1.5 circuits [C], Mars has made almost one and after 780 days the two planets are again in opposition [D].

Mars missions

Mars is a fairly small world and can be properly observed from Earth for only a few months in every alternate year (the average interval between successive oppositions is 780 days). Before the Space Age, man's knowledge of Mars was bound to be incomplete; then in 1962 the Russians made a preliminary attempt to send a space probe past it [1]. They failed because contact with the vehicle was lost at a relatively early stage and was never regained. But in July 1965 the American probe Mariner 4 [2] bypassed Mars at close range and sent back the first detailed information about the planet.

What the Mariner programme revealed

The Martian atmosphere proved tenuous as had been expected in that instead of being made up of nitrogen, with a ground pressure of 85 millibars, it was found to be composed mainly of carbon dioxide, with a pressure of less than 10 millibars, which at once reduced the possibility of any advanced life forms. But the most spectacular discovery was of craters superficially similar to those of the Moon, but with some important differences. The craters were large and some of them had a central peak. This came as a major surprise because scientists had expected that Mars would have a smooth landscape.

The next American probes, Mariners 6 [3] and 7, followed in the summer of 1969, only a few days after Neil Armstrong and Edwin Aldrin in Apollo 11 made their landing on the Moon. They again showed craters, together with mountains, canyons. When measurements were taken there were some more surprises. For instance, the circular white cap known as Hellas, just south of the V-shaped Syrtis Major, proved to be a depression instead of a raised plateau. Hellas is in fact the most depressed area on Mars and it seems to be almost devoid of craters or other important features.

These first three Mariners were fly-by probes, making one pass of Mars and then moving on to solar orbit. Mariners 8 and 9, launched in 1971, were different. They were intended to orbit Mars in order to send back data, including photographs, over a period of months instead of only a few days. Mariner 8 failed immediately after blast-off and fell into the sea, but Mariner 9 [4] was a success, giving us a wealth of new data, including 7,329 photographs.

The craft went into orbit around Mars on 14 November 1971 when almost the whole of the planet was surrounded by dust and little could be seen. Such huge dust-storms are uncommon, although they have been recorded previously in 1909 and 1956, for example. When the dust cleared the spacecraft's cameras discovered volcanoes similar in form to the terrestrial volcanoes of Hawaii, but much larger. For instance, Olympus Mons reaches up to about 25km (15 miles) above the general level of the surface and Arsia Mons to about 20km (12 miles). There were deep rift valleys, and features that looked like dried-out river beds.

Viking's search for life

But did some kind of life await discovery somewhere on this cold, harsh landscape? The United States sent the Vikings to find out. The first swung into orbit on 19 June 1976, the second on 7 August 1976. After a period of photo-reconnaissance to establish

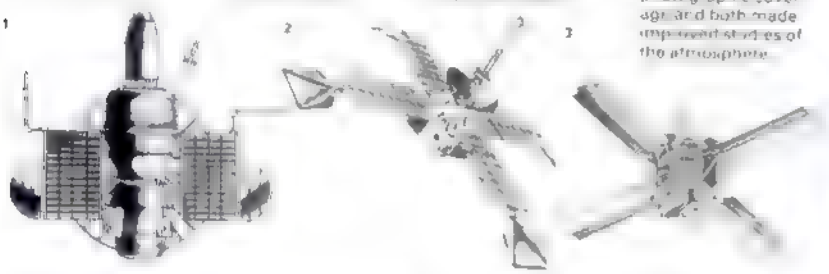
CONNECTIONS

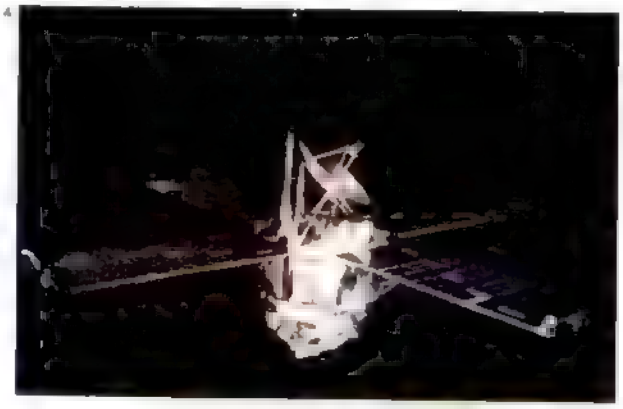
See also

1 Launched by the Russians on 1 November 1962, Mars 1 went into the right orbit, but contact with it was lost early in its flight.

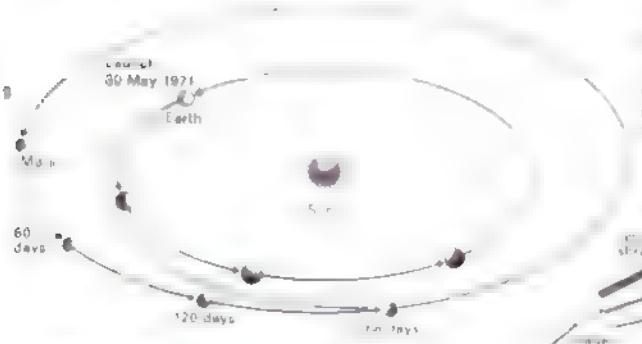
2 The first successful Mars probe, the American Mariner 4, was launched in 1964. It sent back the first pictures of Martian craters.

3 Mariner 6 bypassed Mars in July 1969 and sent back [illegible] Mariner 7 made its pass eight days later. Mariner 6 studied the equatorial regions while Mariner 7's coverage included the south pole. There was extensive photographic coverage and both made improved studies of the atmosphere.

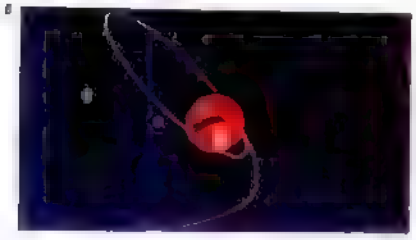




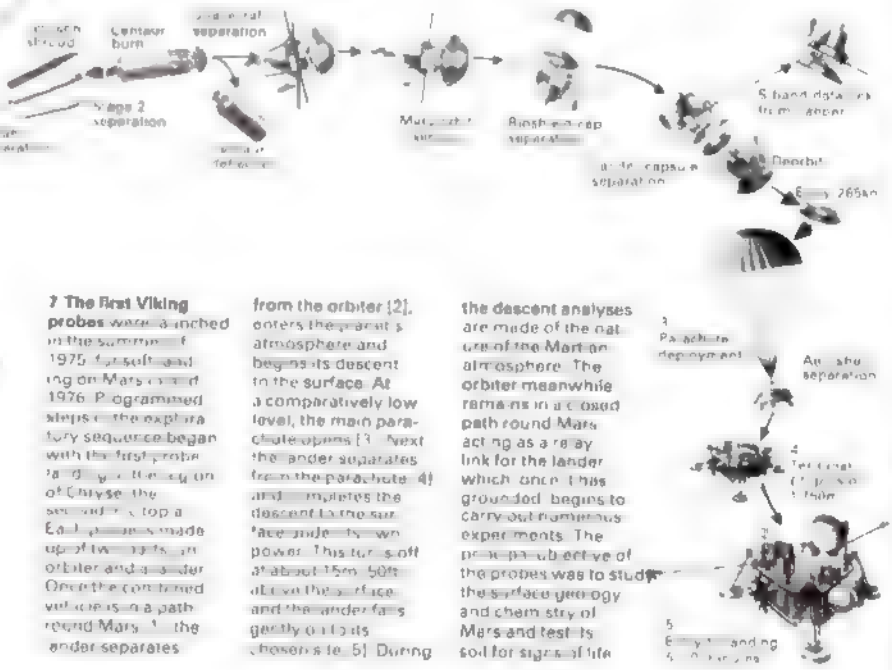
4 Mariner 9, the [illegible] successful Mars probe to date, was a complicated vehicle. The cameras could be controlled from the Earth and their working life exceeded the planned life of the probe. Control was lost only in October 1972 when the gas controlling the attitude jets was exhausted. Mariner 9 measured temperature, analysed the Martian atmosphere and sent back spectacular pictures that enabled astronomers to prepare accurate maps. Both the Martian satellites were also photographed.



5 Mariner 9 reached Mars in November 1971, having started in May. It travelled in a transfer orbit using the Sun's gravity. The probe was accelerated from the Earth and swung outwards in order to rendezvous with its target planet.



6 As Mariner 9 approached Mars, it passed and photographed the two tiny Martian satellites Phobos and Deimos. **Their orbits are shown** together with the elliptical orbit of the probe after it had joined a closed path round the planet.



7 The first Viking probes were launched in the summer of 1975 for soft-landing on Mars in 1976. Programmed steps in the exploratory sequence began with the first probe [illegible] on top of Centaur. The second was a Titan. Each probe was made up of two parts, an orbiter and a lander. Once the combined vehicle was on a path round Mars, the lander separates from the orbiter [2], enters the planet's atmosphere and begins its descent to the surface. At a comparatively low level, the main parachute opens [3]. Next the lander separates from the parachute [4] and completes the descent to the surface under its own power. This turns off at about 15m (50ft) above the surface and the lander falls gently on to its chosen site [5]. During the descent analyses are made of the nature of the Martian atmosphere. The orbiter meanwhile remains in a closed path round Mars acting as a relay link for the lander which, once this grounded, begins to carry out numerous experiments. The principal objective of the probes was to study the surface geology and chemistry of Mars and test its soil for signs of life.

suitable landing sites. Both released landers, that of Viking 1 touched down on the broad Chryse basin in Mars' northern middle latitudes on 20 July. The other landed on 3 September on Utopia Planitia roughly 7,400km (4,600 miles) from Viking 1 and 400km (870 miles) nearer the north pole.

Spectacular photographs received from the landers showed rock-strewn terrain in both places with an overlying reddish soil. The sky was pink from light scattered by red dust particles in the air. Major elements in the soil, as detected by Viking 1's X-ray fluorescence spectrometer, were silicon, iron, calcium, aluminium and titanium.

Both spacecraft dug into the soil under computer control to deliver samples for analysis to the biology laboratories, with results that both surprised and tantalized. The experiments had been designed on the basis that any life forms in the soil must eat and excrete certain basic chemicals.

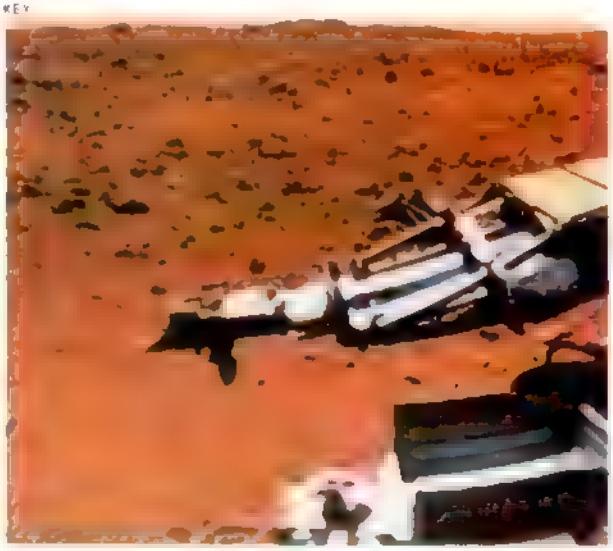
The initial "gas exchange" experiment rapidly detected 15 times as much oxygen as had been expected. Terrestrial life forms with which science is familiar take time to grow and reproduce, and these findings seemed more likely to be the result of a chemical reaction in the iron-rich soil.

Results of the experiments

The result of the "labelled release" experiment also at first looked interesting. If microbes were present, they were expected to take up carbon-14 and give off radioactive wastes such as carbon dioxide, carbon monoxide and carbon methane. A generous amount of carbon dioxide was detected but again the cause seemed to be chemistry.

The "pyrolytic" experiment indicated that something was taking carbon dioxide out of the air in the test chamber and incorporating it into other compounds within the soil, but whether that "something" was biological or chemical was open to doubt.

What worried scientists most was the absence of organic molecules. All three biology experiments showed many signs that could be interpreted as the result of living organisms, only 1 organic compounds were also present. But this did not detract from the success of the programme.

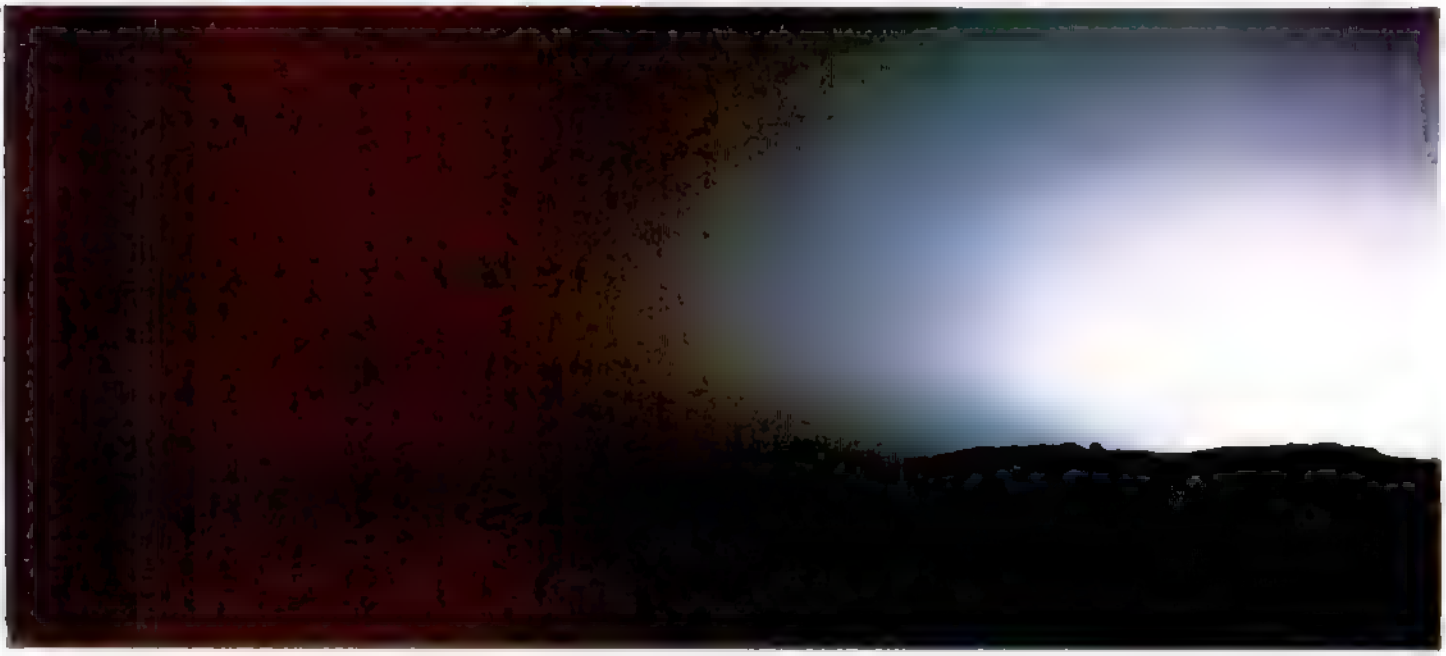


The Sampler scoop of Viking 1 is seen poised over the red orange soil of Chryse Planitia. Some of the rocks are dark and coarse-grained, while others have a lighter, mottled appearance and may have come from lava flows or stream deposits.

8 The first colour photograph of the Martian surface taken by Viking 1 shows orange-red materials covering most of the surface with darker bedrock exposed in patches (lower right). The reddish materials may be limonite (hydrated ferric oxide). Such weathering products form on Earth with water and an oxidizing atmosphere. The sky has a reddish hue, probably due to the scattering and reflection of light from reddish dust particles suspended in the atmosphere. The scene was scanned three times by Viking 1's camera through a different colour filter each time. Colour balance was achieved with the help of a test chart on the spacecraft.

9 A panoramic view from Viking 2 shows the scene on Utopia Planitia, in a composite of three shots. The surface is strewn with rocks out to the horizon, ranging up to several metres in size. Some may have come from the nearby impact crater Mie, which is about 1km (0.62 mile) across. The picture has been electronically rectified to remove the effect of the spacecraft's eight-degree tilt toward the west.

10 This spectacular sunset over Chryse Planitia was photographed by Viking 1. The camera began scanning the scene from the left about 4 minutes after the Sun had dipped below the horizon, and continued for 10 minutes. The Sun had set nearly 3 degrees below the horizon by the time the picture was completed. The Martian surface appears almost black and the horizon line is very sharp.



Mars maps

Most of the Martian features were renamed by Giovanni Schiaparelli (1835–1910) after 1877. His system replaced the older nomenclature; for instance, the "Kaiser Sea" or "Hour Glass Sea" was renamed the Syrtis Major. Further revisions by the International Astronomical Union following the Mariner 9 results have assigned Latin qualifications to those of Schiaparelli's. These are given to topographical as opposed to albedo features (those features associated with the reflecting power of the planet). The new Latin names have been used on the maps below

The western hemisphere

The western hemisphere (below left) includes most of the Acidalia Planitia. To most observers it is the more interesting of the two hemispheres because it also contains some of the greatest of the volcanoes, notably Olympus Mons, which can be seen as a tiny patch from Earth and is surrounded by an extensive, roughly circular area comparatively free from major craters. This area includes Amazonis Planitia to the west, Arcadia Planitia to the northwest and Tharsis Montes to the southeast. The relatively few small craters here were probably produced by meteoritic impact and are not the result of the extensive Martian vulcanism that took place in past ages. Of special significance in volcanic areas are the lava flows, which are particularly well marked around Olympus Mons. Ascraeus Mons, Pavonis Mons and Arsia Mons are also lofty volcanic cones which together make up the Tharsis Montes During the great dust-storm of late 1971, when Mariner 9 reached Mars, these three peaks were some of the few features that could be identified with any certainty because their summits protruded through the top of the dust cloud.

The Vastitas Borealis extends around the north polar region and its southern border is seen across the top of the first map. During the northern winter the white deposits of the polar cap may extend as far south as Tempe Fossae and cover this dark region

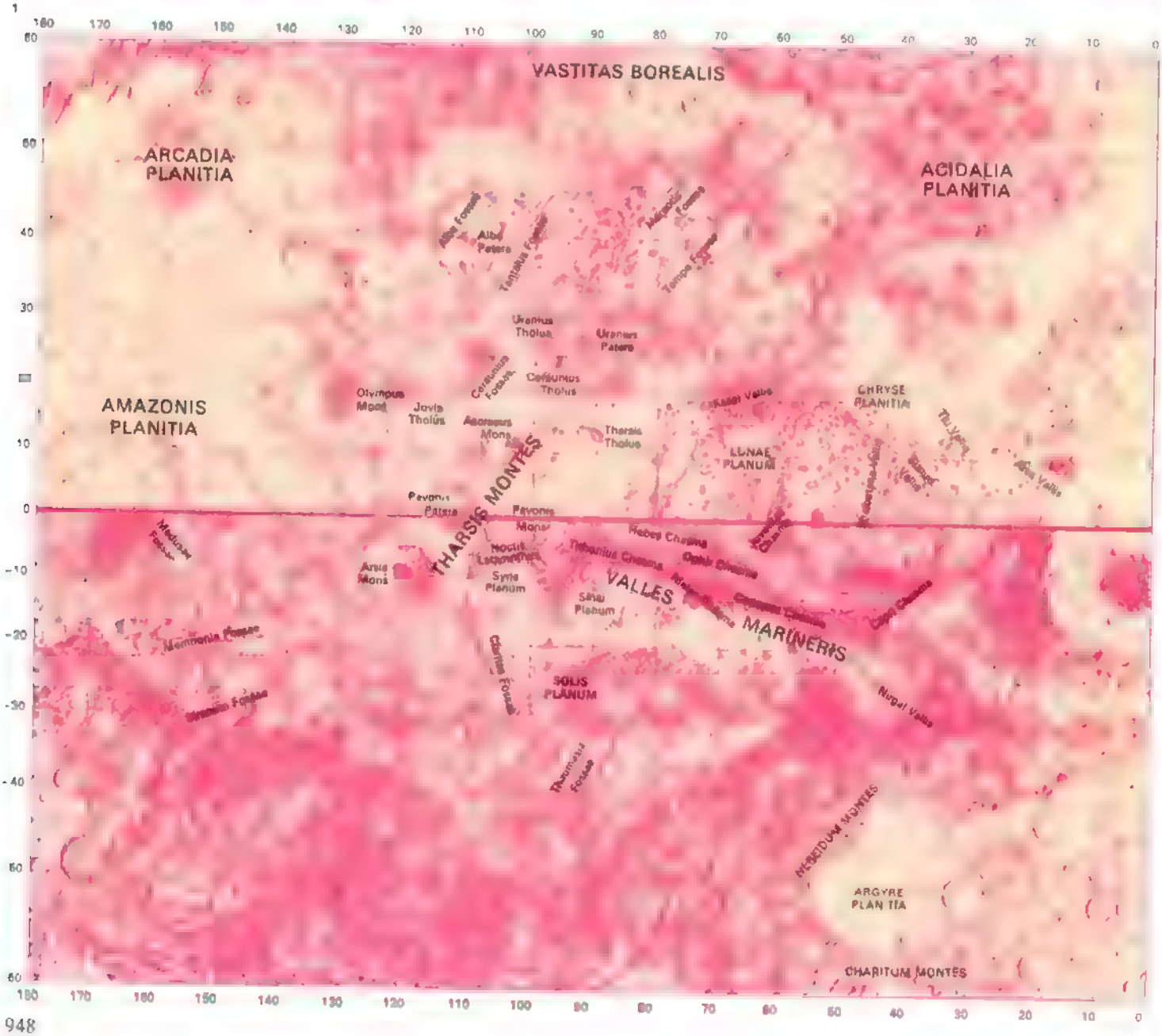
Chryse Planitia was selected as the landing site for the first Viking probe in the summer of 1976. (Viking 2 landed in Utopia Planitia in the eastern hemisphere) About 5° south of the equator and on longitude zero lies the Sinus Meridiani, which used to mark the zero for Martian longitudes, and at a point 20° west and 20° south of that is the dark patch of Margaritifer Sinus. These names did not survive the IAU revisions (and are not on the map) because the markings do not correspond with any obvious topographical features

The hemisphere is dominated by the tremendous system of rift valleys extending eastwards from the Tharsis Montes through Tithonius Chasma, Melas Chasma and Coprates Chasma through Simud Vallis. Immediately south of Tithonius Chasma lies Solis Planum, one of the most variable areas on Mars. Observers since 1877 have noted pronounced changes of shape and intensity Such variations were easy to explain with the old vegetation hypothesis, but now that the dark areas are believed to be inorganic the changes are more puzzling.

The prominent dark patches around the furrows of Sirenum Fossae and the serpentine Nirgal Vallis are two of the classical mare that have disappeared in the new nomencla-

CONNECTIONS

See also



ture. Both are heavily cratered. Russia's Mars 3 probe landed south of Sirenum Fossae in December 1971 but it transmitted for only 20 seconds after arrival and nothing was learned from it.

The eastern hemisphere

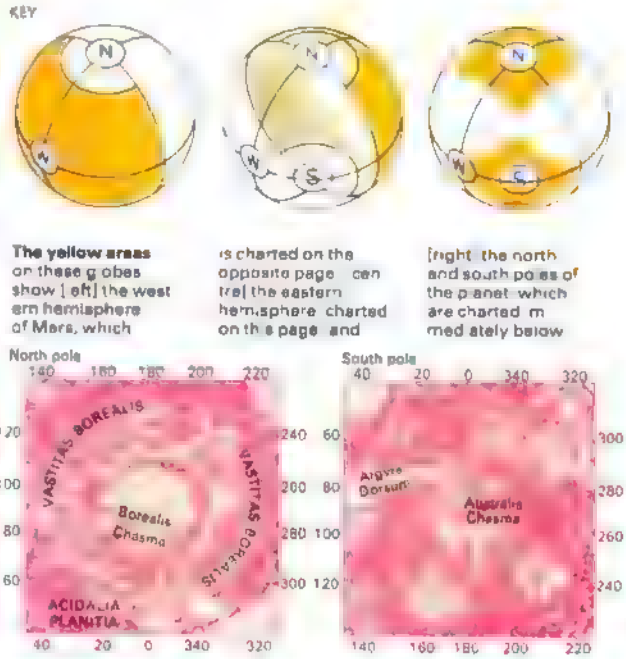
The main feature in the eastern hemisphere (below right) is the Syrtis Planitia, which was recorded by Christiaan Huygens (1629–95) in 1659 and is the most prominent dark marking on Mars. Mariner 9 showed it to be a relatively smooth plateau sloping eastward towards the basin of Isidis Planitia and not, as had been believed, a sunken seabed. Surprisingly, there is little, apart from its colour, to distinguish it from the lighter Isidis Planitia and therefore the conclusion can be drawn alongside it that its prominence is simply due to the low albedo (light reflecting power) of its rocks.

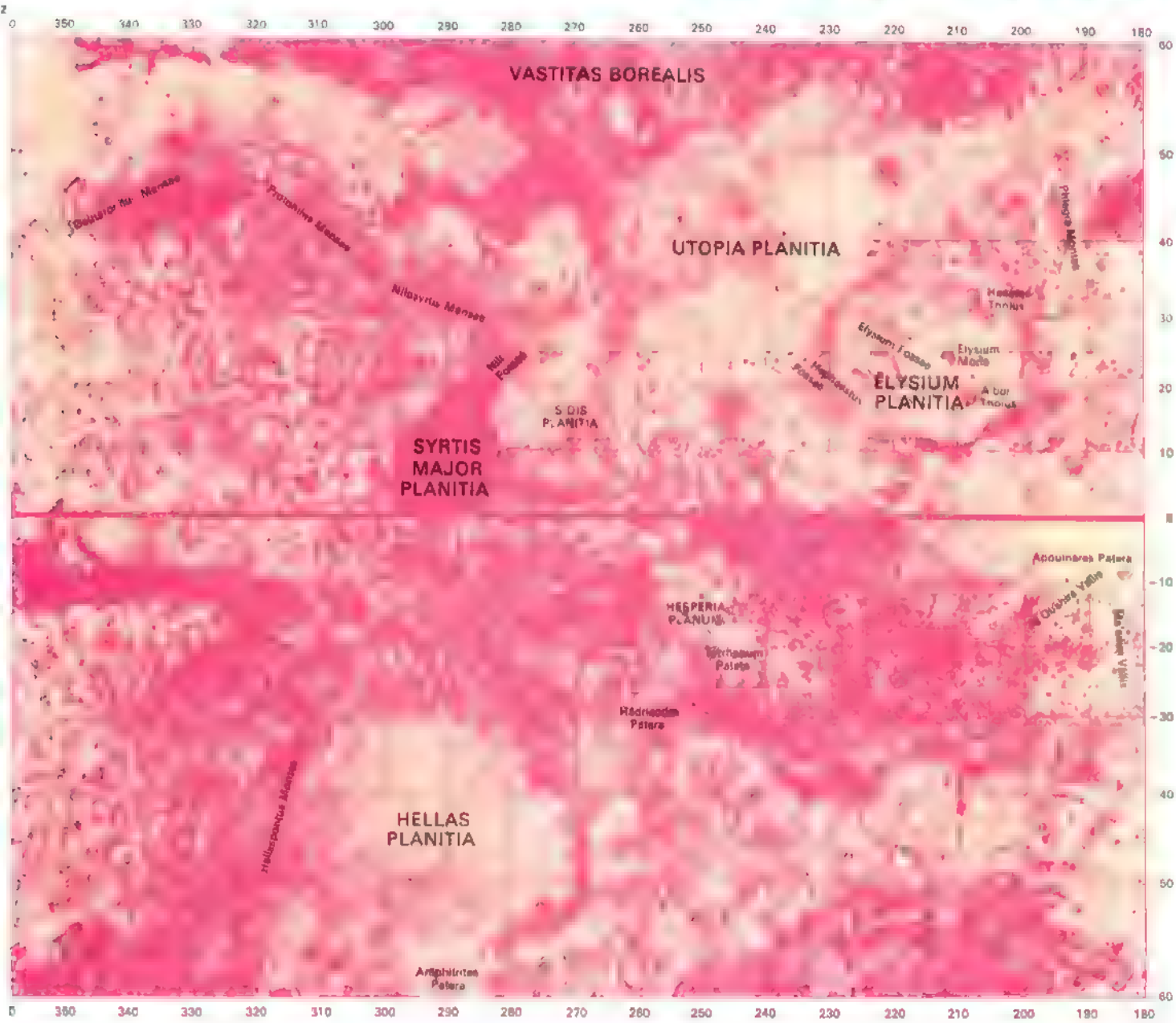
Elysium Planitia to the east of the map is a volcanic province of intermediate geological age. It contains two large volcanic craters (calderas), as well as a clearly marked dome. The dark region north of latitude 55° is the other half of Vastitas Borealis. It may be partly responsible for the dark peripheral region of the polar cap formerly attributed to the visible effects of melting ice. The southern portion of this hemisphere is dominated by Hellas Planitia, which may appear extremely brilliant from Earth and can sometimes be mistaken for an extra polar cap. East of Hellas lie the two prominent dark features known as Mare Tyrrhenum and Mare Cimmerium respectively.

The polar caps

In the north polar region [Key] the white cap never vanishes completely. Part of the Acidalia Planitia is shown, this is the most prominent of the dark features in the northern hemisphere of the planet.

In the south polar zone [Key] the area within 10° of the actual south pole is seen to be smooth and laminated with the summer remnant of the polar cap offset at longitude 45°. The dark surrounding areas are heavily cratered with a prominent ridge, Argyre Dorsum. During the southern winter the polar deposit covers almost the whole area.





Mars panorama

Although Mars is almost certainly about the same age as the Earth (about 4,700 million years), it is so much smaller and less massive that it has evolved more quickly. This fact suggested that the surface features were likely to be more worn and eroded, because although the atmosphere of the planet is thin it is far from negligible.

Theories old and new

It is interesting to look back at what astronomers thought about Mars before 1965, when the first successful probe, Mariner 4, sent back data from close range. It was believed that the dark areas were depressions, probably old sea-beds, while bright regions such as Hellas Planitia and Argyre Planitia were plateaus, also the surface was thought to be gentle in relief with no lofty mountains or deep valleys anywhere. The reality was very different. The first pictures sent back from Mariner 4 showed signs of craters [Key] and, as the probe approached the planet, the photographs became clearer and the general nature of the landscape was no longer in doubt. Instead of being a world with a level landscape, Mars proved to be extremely rough. Yet even from the Mariner 4 evidence it was clear that there were marked differences over various areas and that the surface of Mars was likely to be much more variegated than that of the Moon. Mariner 4 also showed that the atmosphere was much more tenuous than had been previously supposed and the theory of vegetation-filled sea-beds began to look less plausible.

Mariners 6 and 7, in 1969, produced a rather similar picture of Mars and it seemed that there were cratered areas and others that were described as "chaotic", with no particular pattern. Because of technical improvements, pictures were far clearer than those from Mariner 4; much had been learned during the intervening four years.

Mariner 9's discoveries

The most striking discoveries of Mars came in 1971 from Mariner 9, which proved to be a tremendous success – the more so because it also had to compensate for the failure of its predecessor, Mariner 8. After it had approached the planet and photographed the two Martian satellites, Mariner 9 had to wait for dust-storms to subside. It appeared that the dust extended almost to the top of Olympus Mons and Arsia Mons, the highest known points on Mars, with altitudes of more than 20km (12 miles). When the atmosphere became clear, these were revealed as towering volcanoes – and all ideas about the nature of Mars changed yet again. Few astronomers had expected to find volcanoes of a terrestrial type, yet the similarity between Olympus Mons and Arsia Mons on the one hand, and the Earth's Hawaiian volcanoes on the other was unmistakable. The main discrepancy was in scale. Surveys of the height of surface features have been carried out both by radar and by measuring density of the carbon dioxide layer over different areas. As a result it was discovered that the Martian volcanoes are about three times as high as their Hawaiian counterparts and because Mars is a much smaller globe this makes them proportionately even higher.

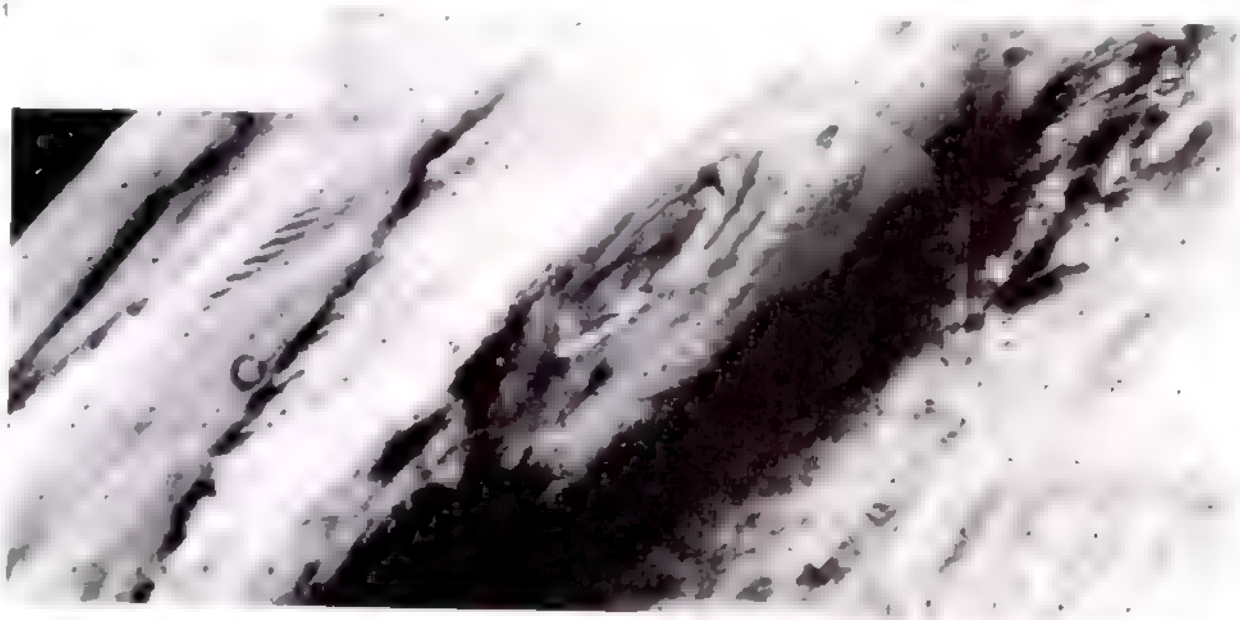
The volcanoes had not been recorded by the earlier Mariners for two reasons. First the times of photography had been very limited

CONNECTIONS

See also

[illegible]

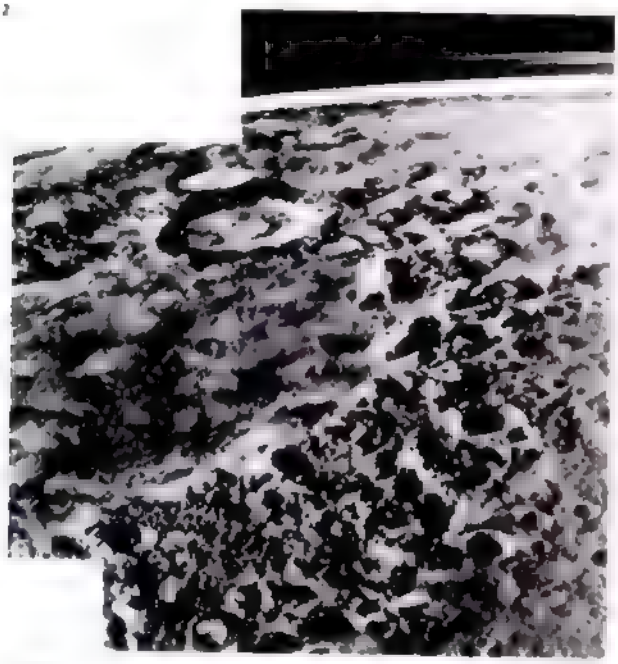
1



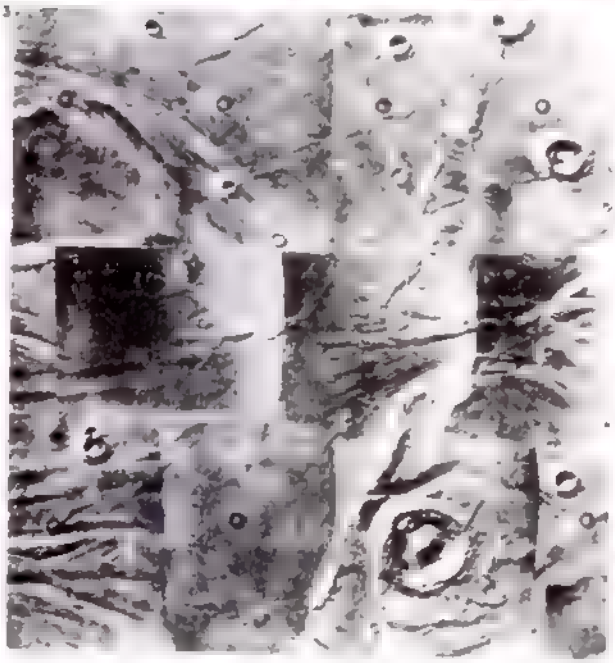
1 The great Martian volcano Olympus Mons was photographed by Viking 1 from 8,000km (5,000 miles) away. The 25km (15.5 miles) high mountain is seen in mid-morning, with clouds that extend up the flanks to an altitude of about 19km (12 miles). The multi-ringed calders, some 70km (43 miles) across, extend into the stratosphere. A well developed wave cloud train stretches several hundred miles behind the mountain.

2 This oblique view across Argyre Planitia extends towards the horizon some 19,000km (12,000 miles) away. Argyre surrounded by heavily cratered terrain – is the relatively smooth plain at left centre.

2



3



3 Eroded channels were photographed near the Viking 1 landing site in Chryse Planitia. They may be the remains of ancient stream beds that fade out near the landing ellipse, suggesting that the area is a sediment basin. The picture is a mosaic of 15 photographs taken by Viking Orbiter 1 from a distance of 1,680km (1,040 miles) and has an area of about 250 by 200km (155 by 120 miles). The lava flows are broken by faults that form ridges and are peppered by meteorite impact craters. A small stream flowed northward (toward upper right) from Lunae Planum, crossed the area and flowed toward the east. In places the water may have ponded behind the ridges before cutting through to form "water gaps".

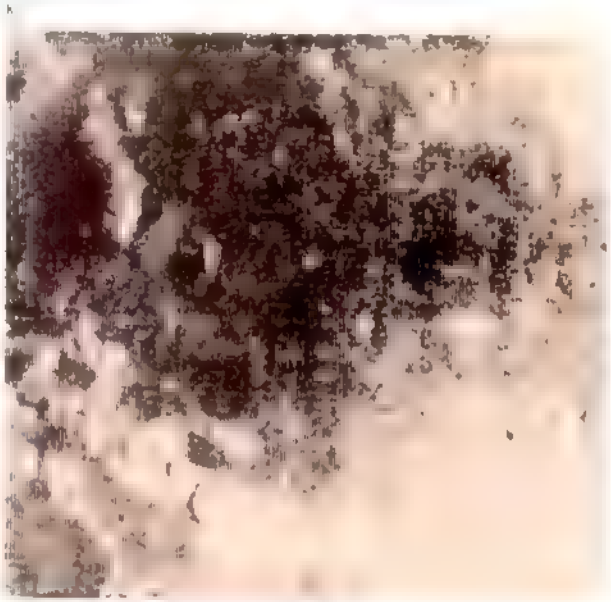
Mariners 4, 6 and 7 each made only one pass of Mars and then moved on into never-ending orbits round the Sun, whereas Mariner 9 and the Vikings were put into orbit round the planet and were capable of sending back data until their power was exhausted. Second, there had been a great improvement in the computerized "cleaning-up" of Mariner pictures, it became possible to make vast improvements by electronic methods and the process took only a short time.

Photographic coverage

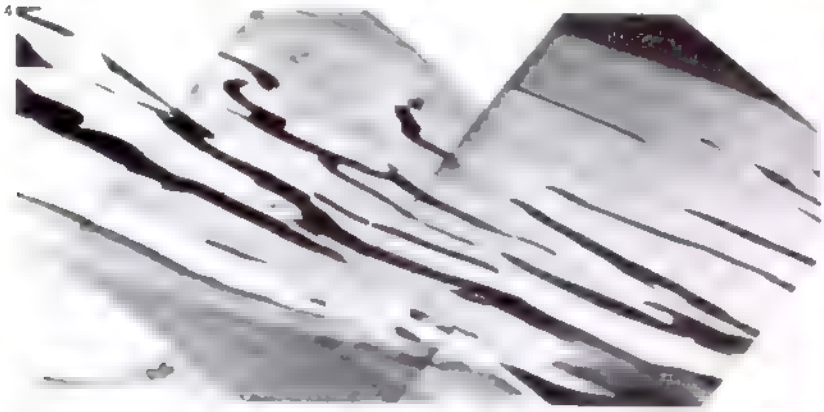
Our knowledge of the Martian surface has been revolutionized by Mariner 9 and the Vikings and a full analysis of their findings will take many years to accomplish. The most striking aspect is the diversity of the features in different areas. Regions that are thickly cratered are succeeded by relatively level areas, apart from great volcanoes there are drainage canyons and deep basins, of which Hellas and Argyre are the best examples. The equatorial canyons Valles Marineris cut deeply into the surface of Mars and stretch for nearly a third of its circumference. In places the canyon walls appear to have been modified by huge landslides, in others by headward erosion to form integrated tributary systems. Elsewhere faulting seems to have predominated.

When the orbit of Viking Orbiter 1 was changed so that the whole planet revolved beneath it, the craft's water vapour and temperature mappers made another major discovery. It was found that the northern polar ice cap is mainly frozen water and not dry ice (frozen CO_2) as most scientists had believed. In some places the ice may be 100 to 1,000 metres (330 to 3,300 feet) thick. Photographs show that it virtually fills some craters.

Over large areas of the planet there is evidence of volcanic activity and water erosion. In Mars' geological past, when the atmosphere was very much thicker than it is today, torrential rains must have flooded the basin areas and cut channels through rock and desert. Some of the drainage channels may even have been caused by geothermal action. Another theory is that they were formed by flowing molten lava.

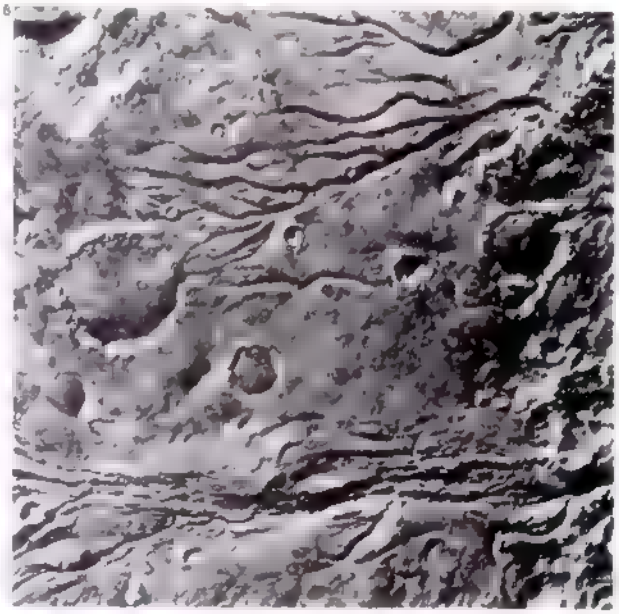


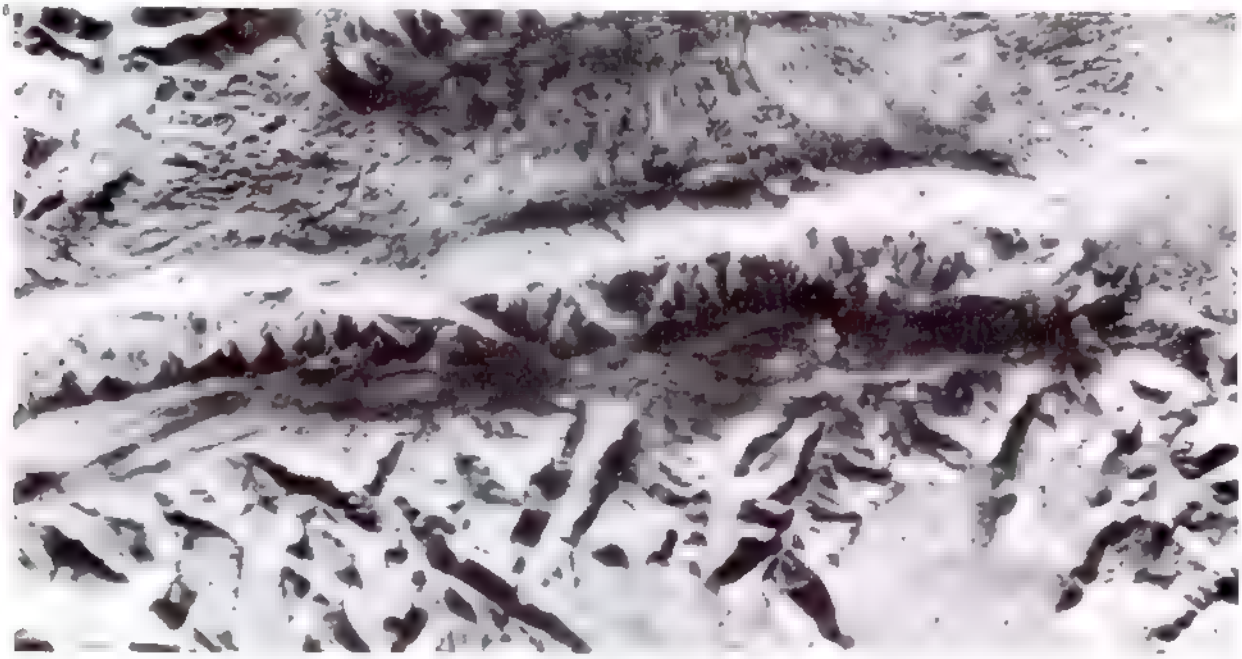
In 1965 Mariner 4 sent back the first close range photographs of Mars. This picture took 8hr 35min to transmit.



4 Large amounts of water still exist on Mars locked up in the north polar cap. In these overlapping pictures taken by Viking Orbiter 2 from 4 000 km (2 480 miles) the northern cap has receded to its smallest size as mid-summer approaches in the northern hemisphere. The solid white area near the top (north) is ice mainly water ice with most of the frozen CO_2 evaporated off. The dark bands are regions devoid of ice.

5 The terrain in this Viking Orbiter 1 picture slopes from west to east with a drop of about 3km (1 86 miles). The channels are a continuation of those to the west of the Viking 1 landing site in Chryse Planitia. They are suggestive of a massive flood of waters from Lunae Planum, across the intervening cratered region, and into the general region from which Viking 1 took its soil samples.





6 Valles Marineris is an enormous equatorial canyon that stretches nearly a third of the way round Mars. The far wall shows several large landslides that probably took place in series and perhaps were triggered by Marsquakes. Along the near wall another widening process seems to have occurred: a series of branch channels cuts into the plateau at the bottom. These may have been formed by slow erosion as a result of the release of ground water, or by mass wasting processes in which rock debris moves downhill as ground ice freezes and thaws. The photomosaic was made from pictures taken by Viking Orbiter 1 from a distance of some 4 200km (2 600 miles).

The moons of Mars

In 1877 Asaph Hall (1829–1907), using the 26in (66cm) Washington refractor in the United States, discovered two satellites of Mars; they were subsequently named Phobos and Deimos. Both are extremely small and are in no way comparable to the Moon. They had not been found before 1877, despite periodical searches, because they are both extremely faint.

Phobos and Deimos before Mariner

Telescopically, Phobos and Deimos appear as small, star-like points, but they caused a great deal of interest in the pre-Space Age period because of their unusual orbits [1]. Phobos moves round Mars at a mean distance of only 9,350km (5,800 miles) from the centre of the planet, so that the distance between Phobos and the Martian surface is about the same as that between London and Aden. The revolution period is only 7hr 39min, and since the rotation period of Mars is 24hr 37min, the "month", reckoning by Phobos, is shorter than the Martian day. In relation to Mars, Phobos rises in a westerly direction and sets toward the east; it is above the horizon for only 4.5 hours at a time during which it goes through more than half its cycle of phases and the interval between successive risings is a little more than 11 hours. The apparent diameter never exceeds 12.3 degrees, less than half that of the Moon as seen from Earth, and the amount of light sent down to the Martian surface is about the same as Venus sends to Earth. Phobos transits the Martian view of the Sun 1,300 times each year, taking 19 seconds or so to pass across the solar disc.

Even when above the Martian horizon Phobos remains eclipsed by the shadow of the planet for long periods and from Martian latitudes greater than 69 degrees it never rises at all. The orbit of Phobos is practically circular and the inclination of the orbit to the equatorial plane of Mars is only a little more than one degree.

Deimos, smaller and more remote (23,500km [14,600 miles] from the centre of Mars), has a revolution period of 30hr 14min and remains above the Martian horizon for 2.5 days consecutively, but it sends less light to Mars than Sirius sends to Earth and to an observer on Mars its phases would be almost imperceptible. Its maximum diameter is only about 12km (7.5 miles).

The nature of the two satellites is a matter for debate. They could be ex-asteroids that have been captured from the main minor planet zone. Some years ago it was suggested that Phobos was spiralling slowly down towards Mars and would collide with the planet in the foreseeable future, this led to a remarkable suggestion that it was being "braked" by the very tenuous Martian atmosphere. As its mass would have to be almost negligible for that to happen, the conclusion was reached that Phobos was a hollow space station built by the Martians. The idea came from Joseph Shklovsky (1916–), an eminent Soviet astronomer but it received little support.

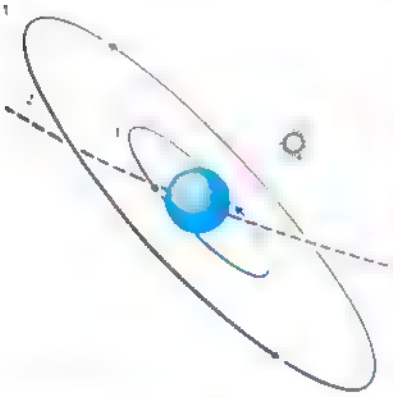
Mariner 9 discoveries

The first positive information about the satellites came from Mariner 9, which approached Mars in late 1971 and entered an orbit round the planet. During its approach Mariner photographed both Phobos and Deimos and

CONNECTIONS

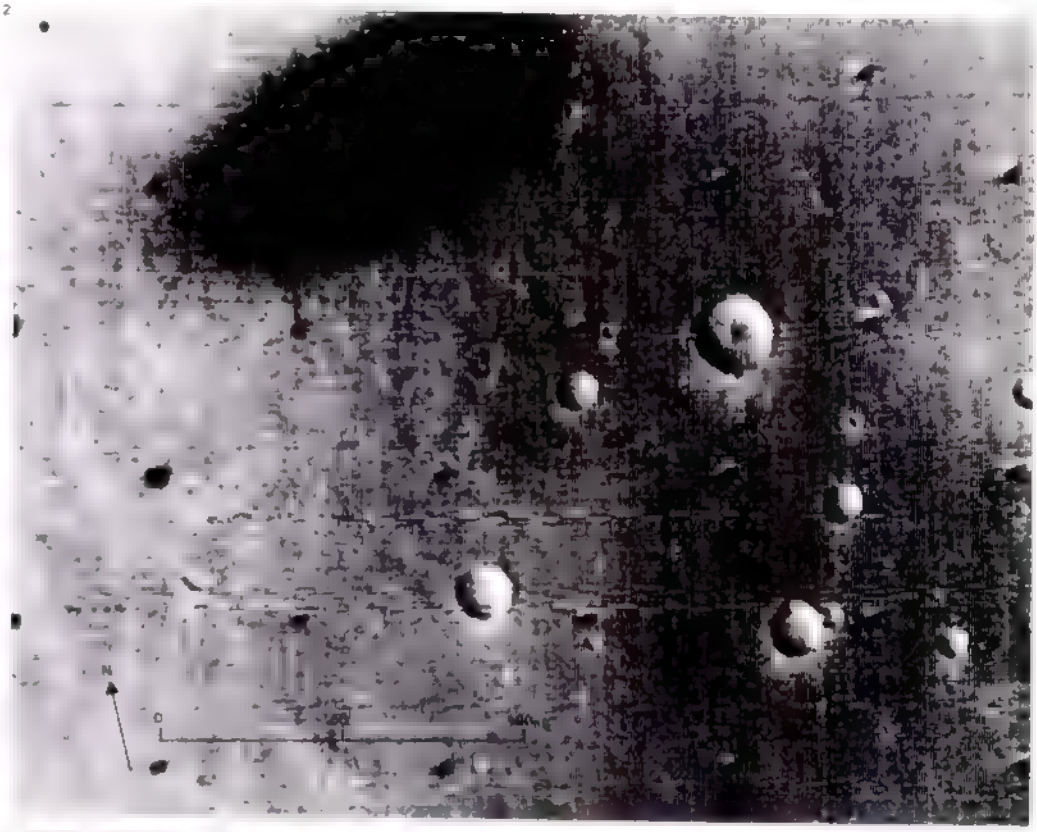
See also

1



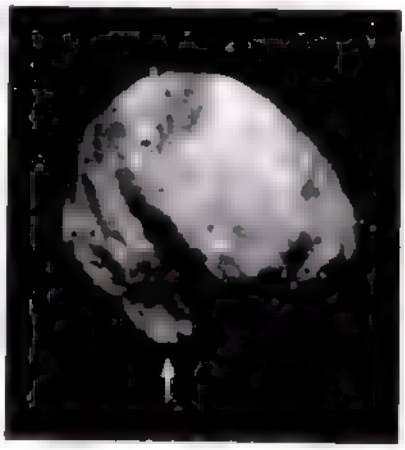
1 Both the satellites of Mars move in orbits that are practically circular and in the plane of the planet's equator. Phobos [1] is remarkably close to Mars and may approach to within 5,800 km (3,600 miles). It is the only known natural satellite with a revolution period shorter than the rotation period of its primary and as seen from the planet it rises in the west and sets in the east. Deimos [2] is much farther out, with a period of 30hr 14min.

2



2 The Aethiopis region on Mars (lat. 14° N, long. 235°) is shown in detail in this Mariner photograph. The elliptical dark patch is the shadow of Phobos, measuring 130km (80 miles) by 280km (175 miles), so that an observer standing in the shadowed area would see a transit of Phobos across the Sun's disc. Because Phobos seems so much smaller than the Sun as seen from Mars, it could not produce a total eclipse.

3 A



3 This series of photographs of Phobos (A–C) was taken by Mariner 9 in 1971. The approximate position of the south pole of the satellite is indicated by the arrow in each photograph. There is a pronounced surface bulge towards the top: this is the "synchronous" bulge permanently turned towards Mars, as the rotation period of Phobos is exactly the same as the time that it takes for the satellite to complete one journey round the planet – 7 hours 39 minutes.

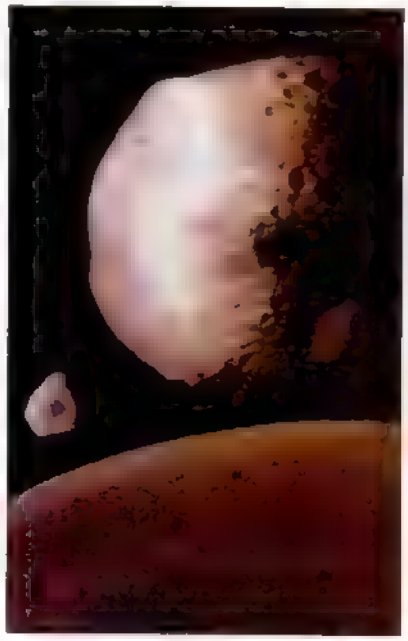
showed that both are irregular in form. Phobos proved to be shaped rather like a potato, with a longest diameter of 28km (17 miles) and a shortest diameter of 20km (12 miles). Its surface is pitted with craters, of which the largest formation (since named Stickney) has a diameter of 6.5km (4 miles) [4]. More than 50 features have now been charted [5], of which seven have been given official names: Roche, Wendell, Todd, Sharpless, D'Arrest, Stickney itself and the Kepler Ridge. Variations in surface height amount to as much as 20 per cent of the satellite's radius. The rotation is synchronous, so that Phobos always keeps far-side face to Mars, the longest axis points to the planet.

Mystery craters

The origin of the craters of Phobos is not definitely known. Meteoritic impact has been suggested; the Japanese astronomer S. Miyamoto prefers the theory that the craters are of the big whole variety, produced during the cooling period of the satellite. If impact is the cause, it must be agreed that Phobos has been severely battered: the diameter of Stickney is approximately one-quarter the diameter of Phobos.

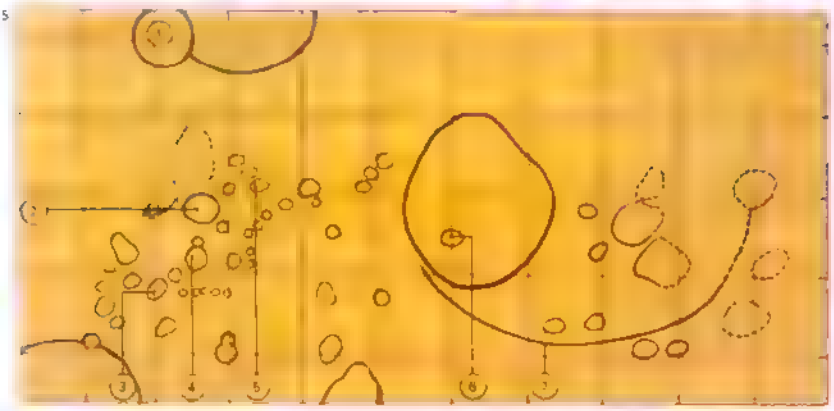
When Viking Orbiter 2 swung within 880km (545 miles) of Phobos in September 1976, it was able to photograph objects on the surface down to 40m (130ft) across. The features observed indicated that Phobos has the structural strength of solid rock, the main material is probably basalt. The escape velocity is a mere 20kmh (12mph), so that there can be no trace of atmosphere. The tiny moon was seen to be heavily cratered as expected but, surprisingly, showed striations (grooves) and chains of small craters. Similar crater chains appear on Earth's Moon, Mars and Mercury were formed by secondary cratering from a larger impact but craters are not so easily explained in the case of a small low gravity body.

Deimos is of the same general type as Phobos, although smaller. It too is cratered, and the two main formations have been named Swift and Voltaire, after two writers who predicted in the eighteenth century that Mars would eventually be found to possess two satellites.



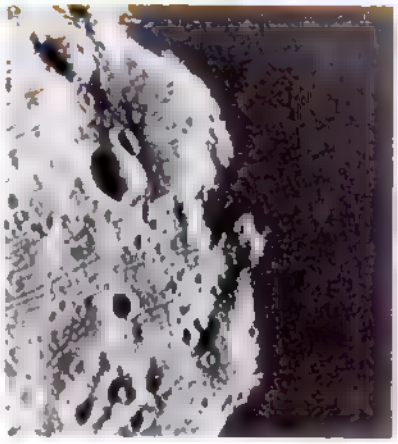
Phobos and Deimos the satellites of Mars, are much smaller than the Moon. Here they are shown as seen from the Martian equator while the Moon is shown as seen from Earth, all three [illegible] scale. The nature of the satellites remains uncertain, but the Mariner 9 photographs suggest that both Phobos and Deimos are solid rock. They are quite different from our own Moon: it is possible that they are captured asteroids. Neither provides as much illumination as the Moon at night. Phobos gives about as much light to Mars as Venus does to Earth, while Deimos is still more faint. The surface of both satellites is exceptionally dark.

4 These Mariner 9 photographs of Phobos are the clearest pictures of the satellite so far obtained. Again, the arrow indicates the approximate position of the south pole. The largest crater, 6.5km (4 miles) in diameter, is Stickney. The apparent indentation (upper right) in pictures A and B shows up as a crater in the third (C). Phobos is not even approximately spherical: its form is quite irregular.



5 This Phobos map was compiled from the Mariner 9 photographs. Fifty craters have been charted, of which six have been given official names: Roche, 1; Todd, 2; D'Arrest, 3; Sharpless, 4; Wendell, 5; and Stickney, 6. The ridge, 7, has been named the Kepler Ridge. There is no doubt that the whole surface of Phobos is heavily cratered and fractured, which is probably the result of large meteoritic impacts, although this is not certain.

6 This closest ever view of Phobos may help to resolve the question of how the tiny moons of Mars originated. Clues are a crater with a central peak, crater chains indicated by arrows running parallel to the equator, and striations (grooves) covering more than half the surface. One theory supposed that the grooves were caused by Phobos passing through a swarm of smaller bodies. The smallest visible object is 40 metres (130 ft) in diameter.



7 Deimos was viewed by Viking Orbiter 1 from a distance of 3,300km (2,050 miles). Mars is to the left. About half the side facing the camera is illuminated and the lighted portion measures about 12 by 8km (7.5 by 5 miles). While Mariner 9 photographs taken from greater distances showed only a few large craters, at least a dozen are prominent here, the largest having diameters of 1.3km (0.8 miles) and 1km (0.62 miles). A linear feature appears near the top.

Minor planets

The Solar System is divided into two main parts by a wide gap between the outermost of the inner planets, Mars, and the first of the giants, Jupiter. A numerical relationship known as Bode's law, discovered by Titius of Wittenberg (1729–96), but popularized by Johann Bode (1747–1826) in 1772, led astronomers to suppose that there might be an extra planet there.

At the end of the eighteenth century, a group of astronomers headed by Johann Schröter (1746–1816) and Baron von Zach (1754–1832) formed themselves into what became known as the "celestial police" with the express intention of searching for the new planet. In fact they were forestalled.

New discoveries: the minor planets

On 1 January 1801, Giuseppe Piazzi (1746–1826), at Palermo in Sicily, discovered a star-like body that moved perceptibly from night to night and that proved to be a planet moving in the gap between Mars and Jupiter. It was named Ceres, in honour of the patron goddess of Sicily.

During the next few years the "celestial police" found three more planets: Pallas, Juno and Vesta. Together with Ceres, they became known as the minor planets, asteroids or planetoids. All are small and, apart from Ceres, less than 500km (300 miles) in diameter. Only Vesta is ever visible to the naked eye.

No more asteroids were evident and the "police" disbanded, but in 1845 a fifth asteroid, Astraea, was discovered by Karl Hencke (1793–1866) and since 1850 no year has passed without further discoveries: the swarm may well exceed 50,000.

In 1977 a faint object (19th magnitude), was detected travelling between Saturn and Uranus at a mean distance of 1400 million miles. Believed to be an exceptional asteroid—of probably 1000km (600 miles) diameter—it has been named Chiron and, it has been suggested, may be an ex-satellite of Saturn.

Irregular orbiters

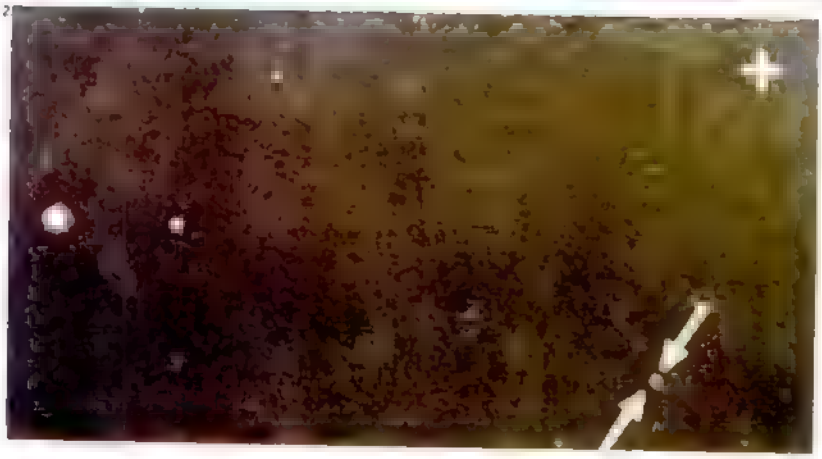
Not all the asteroids keep strictly to their main zone [4]. In 1898 Carl Witt, at Copenhagen, discovered number 433, Eros, which can move well inside the orbit of Mars and occasionally approaches to within 24 million kilometres (15 million miles) of the Earth, that happened in 1931 and again in 1975. In 1931 Eros was extensively studied because exact calculation of its orbit can help in measuring the length of the astronomical unit or Earth–Sun distance. In shape Eros is elongated, with a long diameter of about 27 km (17 miles) and a short diameter of less than 16km (10 miles). Although small, Eros is larger than other so-called "Earth grazers" such as Hermes, a mere 1km (0.6 miles) or so in diameter, which brushed past the Earth in 1937 only 780,000km (485,000 miles) away, less than twice the distance of the Moon. The Earth could be hit by such an asteroid and much damage would result, although the chances of a direct collision are extremely slight.

One asteroid, Icarus, actually approaches the Sun closer than does the planet Mercury. It must undergo some of the most extreme temperature changes in the Solar System. At the closest point in its orbit, only 28 million kilometres (17 million miles) from the Sun, the surface temperature on Icarus must be

CONNECTIONS

See also

1 This photograph taken by Max Wolf (1863–1932) shows a star field together with two streaks that represent asteroid trails. The photograph is a time exposure, during it the driving mechanism of the telescope was adjusted so as to follow the stars (to compensate for the rotation of the Earth). The stars remained in the same relative positions while the asteroids shifted perceptibly against the background. Wolf was the great pioneer of this method of asteroid discovery. Previously the method had been to chart the same area of the sky for several consecutive nights, so that any star-like object that moved could be identified as an asteroid. Wolf's photographic method was far quicker and more efficient. Discovering asteroids is now very much easier, but keeping track of them and computing their orbits is time consuming.



2 Vesta, discovered in the early nineteenth century and shown here between the arrows, is the brightest but not the largest of the minor planets. It looks exactly like a star although its night to night movements betray its true nature. The cross to the upper right shows Vesta's position 24 hours later. Photographs do not show surface detail—asteroids are too small. Their rotation periods can be determined by variations in brilliancy.

3 Icarus, about 1.5 km (1 mile) in diameter, was discovered by Walter Baade (1893–1960) in 1949. The orbit of Icarus (1) is highly inclined at 23°. It is the only asteroid that is known to have its perihelion inside the orbit of Mercury (2).

more that 300 (906 E) at its aphelion (farthest point) only 200 days later, it has moved out to a distance of 295 million kilometres (183 million miles) – well beyond that farthest orbital point of Mars.

On the other hand, number 944, Hidalgo, has an eccentric path that takes it out almost as far as the orbit of Saturn while the members of the Trojan group move in the same orbit as Jupiter. One group keeps approximately 60 degrees ahead of Jupiter and the other group 60 degrees behind. There is no danger of a collision. Although the Trojans are large by asteroid standards, their distance from Earth makes them faint objects.

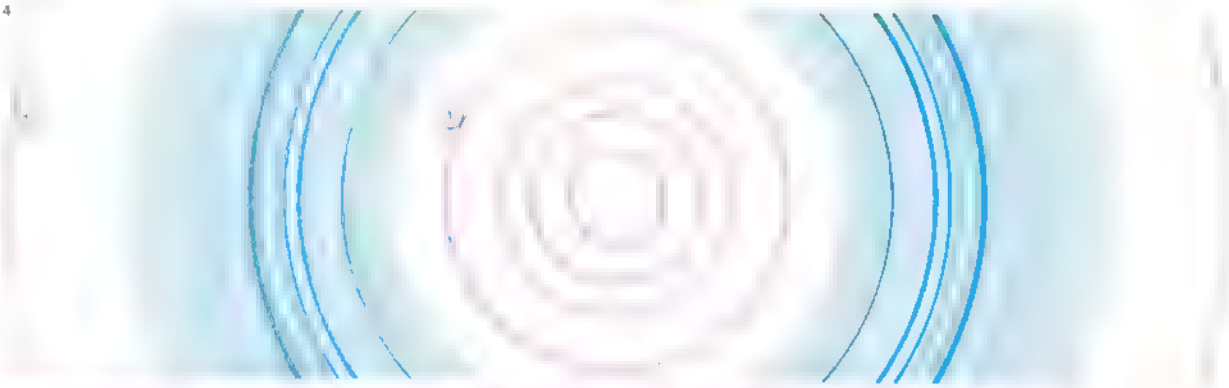
Through a telescope asteroids look exactly like stars and the only way to identify them is by checking their movements from night to night. Modern discoveries are made photographically. During a time exposure, an asteroid will often move enough to leave a trail on the plate rather than a point of light. Asteroids can therefore be a nuisance to astronomers. Photographic plates exposed for quite different reasons are often found to be covered with asteroids and each has to be individually identified, which wastes time.

It is not yet known what asteroids are made of, but Mariner 9 photographs of the two dwarf satellites of Mars (Phobos and Deimos), which may well be captured asteroids, suggest that the surfaces of many may be pitted with craters. Some of the smaller satellites of the giant planets – the outer members of Jupiter's family, Phoebe in Saturn's and Nereid in Neptune's – also may be captured asteroids.

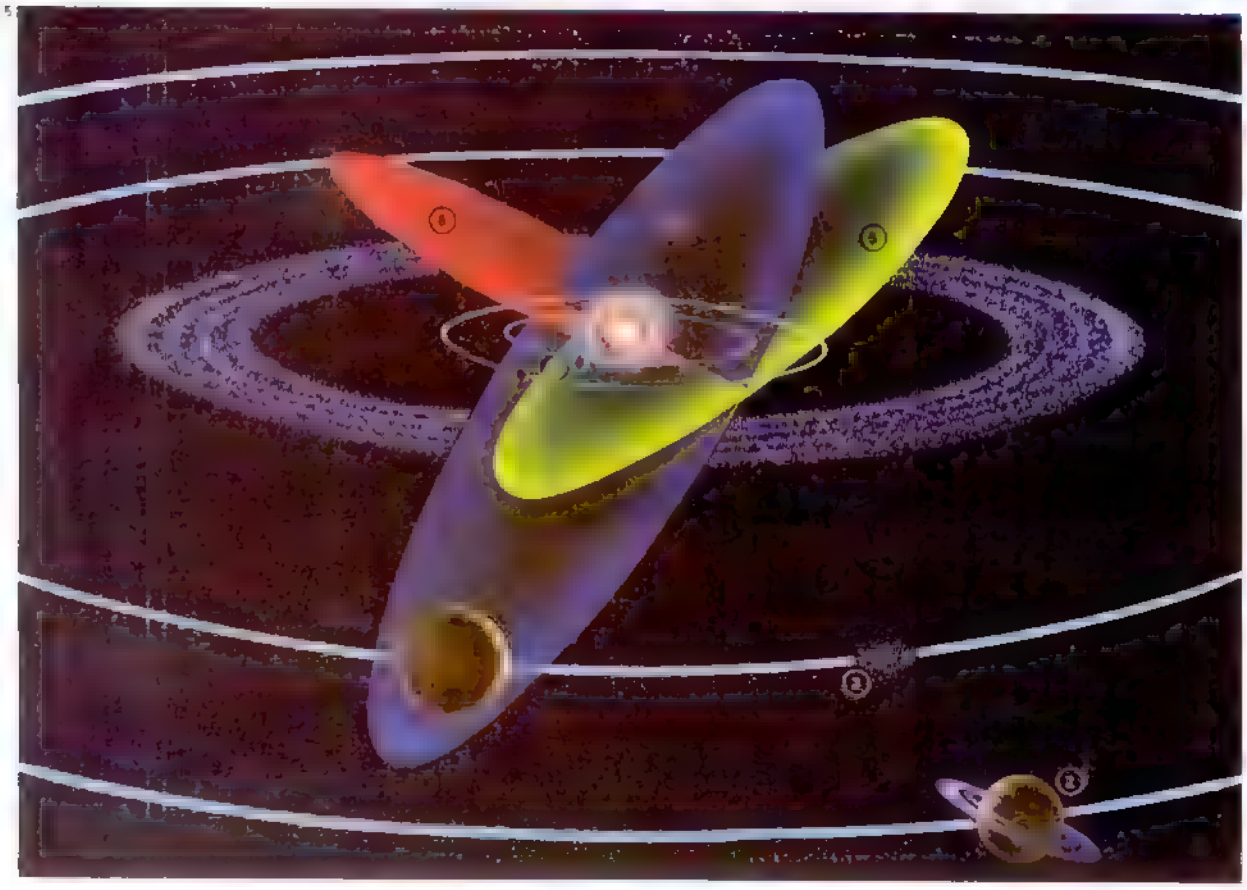
Origin of asteroids

The origin of the asteroids is still uncertain. According to one theory, they represent the debris of a former planet (or planets) that used to orbit the Sun beyond the path of Mars and met with some disaster in the remote past; but on the whole it seems more likely that they never formed a large body. The immensely powerful gravitational effect of Jupiter could prevent a large planet from forming in the region of the asteroid zone. It is also worth noting that all the asteroids combined would still not make up one body as large or as massive as the Moon.

The sizes of the first four asteroids to be discovered. Ceres (C), Vesta (D), Pallas (E) and Juno (F), together with the irregularly shaped Eros (B), are here compared with the Moon (A). Their diameters are difficult to measure, being so small. Earlier measurements of Ceres gave 685km (426 miles) but new methods show that it is much larger – between 1,000 and 1,200km (600 and 750 miles).



4 Most of the orbits of the minor planets lie well beyond the orbit of Mars (1) and well inside that of Jupiter (2). With all the distribution there are certain regions, known as Kirkwood gaps, in which there are fewer asteroids. These gaps lie at particular orbital distances from the Sun where Jupiter's gravitational field has forced the asteroids into different orbits.



5 The orbits of the planets are shown from Earth (1) out to Saturn (2), together with those of some asteroids which are of particular interest (the illustration is not to scale). Most of the asteroids move in the region between the orbits of Mars and Jupiter. The Trojan asteroids (3) move in the same orbit as Jupiter but keep their distance and collisions are unlikely to occur. One group moves 60 degrees ahead of the planet and the other 60 degrees behind, although they move around for some distance to either side of their mean positions. Hidalgo (4) has a path which is highly inclined and so eccentric, much like a comet, that its aphelion is not far from the orbit of Saturn. Amor (5) and Apollo (6) belong to the so-called Earth-grazing asteroid group. All the Earth-grazers are very small. Amor has a diameter of 8km (5 miles) and Apollo about 2km (1.25 miles).

The planet Jupiter

Beyond the asteroid belt lies Jupiter, the giant of the planets. Jupiter is more massive than all the other planets together, and it has a diameter about 11 times that of Earth. It orbits the Sun at an average distance of 778,300,000km (483,600,000 miles), taking 11.86 years to complete a revolution.

Jupiter is well placed for observation for several months in each year: it appears to the naked eye as a brilliant "star". A small telescope will show the planet's disc, and reveal that it is not circular but bulges noticeably at the equator. This is a result of Jupiter's rapid spin. Although Jupiter is such a large planet, its rotation period ('day') is less than ten hours long. The equatorial diameter is 143,000km (89,000 miles), noticeably larger than the polar diameter of 135,000km (84,000 miles).

Interior structure

Jupiter's mass is 318 times that of the Earth, and because this mass fills a volume 1,400 times greater than that of Earth, Jupiter's density must be much lower than the Earth's. Thus the giant planet cannot be composed of rock, and since the 1930s astronomers have known that Jupiter is made up of mainly hydrogen and helium. This has been confirmed by subsequent Earth-based observations, and by the four space probes that have flown past Jupiter: Pioneer 10 in 1973, Pioneer 11 a year later, and Voyagers 1 and 2 in 1979.

Unlike the inner planets, Jupiter does not have a solid surface. At its centre there may be a rocky core considerably more massive than the Earth, but the vast bulk of the planet consists of a hydrogen and helium mixture which is so compressed that it behaves more like a liquid than a gas in the lower regions. Deep within the "liquid" bulk of Jupiter the hydrogen bears some of the properties of a metal - in particular, it is a very good conductor of electricity. This makes it a very efficient electromagnet, and electric currents in Jupiter's core produce a magnetic field 20,000 times stronger than that of Earth.

Astronomers have known of Jupiter's magnetism since 1955, when B.F. Burke and W. Franklin detected strong radio emission from the planet, and the field has now been measured by the four probes which have flown through its magnetosphere. The probes have found that the field traps very energetic electrically-charged particles – as do the "Van Allen belts" around the Earth.

Jupiter's atmosphere

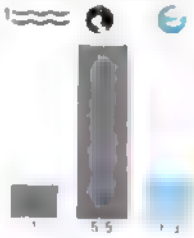
The outermost layer of Jupiter is gaseous in nature. Earth-based and space probe measurements show that it is about 89 per cent hydrogen and some 11 per cent helium. Other molecules total only 0.2 per cent: they are, in decreasing order of abundance, water vapour, methane, ammonia, neon and hydrogen sulphide.

The visible "surface" of Jupiter consists of gaudily-coloured cloud layers floating in this atmosphere. The highest clouds are white, and probably composed of small crystals of frozen ammonia. Between them, we can see down to deeper layers of cloud, which may consist of ammonium hydrosulphide. These clouds display a whole range of brown, orange and yellow colours, no doubt caused by traces of other, highly-coloured, sulphur compounds.

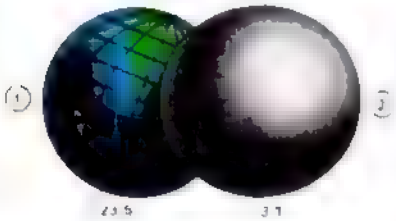
CONNECTIONS

See also

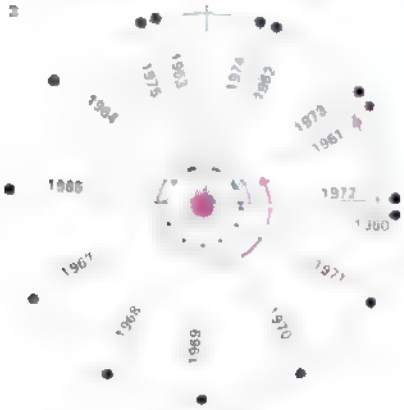
The Moons of
Members of the Solar System
Exploring the planets



1 Jupiter's density is only 1.3 times that of water whereas the Earth's is 5.5. However, this is the mean density of the whole of Jupiter's globe. The outer layers are relatively tenuous while the core is much denser.



2 The Earth's axis is tilted at an angle of 23.5 degrees from the perpendicular to the plane of the orbit (1); in the case of Jupiter (2) the tilt is only 3.1 degrees. In fact Jupiter is practically "upright". Of the other principal planets, only Mercury's axial inclination is like this.



3 Jupiter comes to opposition at mean intervals of 399 days, so that astronomers can observe it clearly for several months each year. The diagram shows the opposition positions between 1960 and 1975. Because Jupiter's orbit, like that of the Earth, is eccentric, the opposition distance ranges from 589-669 million km (366-416 million miles). In 1975 Jupiter was at its closest to Earth.

4 The surface of Jupiter shows bright zones and dark belts. The nomenclature always used by observers is given here (south is at the top): [1] South Polar Zone, [2] South South Temperate Belt, [3] South Temperate Belt, [4] and [5] South Equatorial Belt, frequently seen to be divided into two well-marked components, [6] Equatorial Band, [7] and [8] North Equatorial Belt, also frequently divided into two components, [9] North Temperate Belt, [10] North North Temperate Belt, [11] North North North Temperate Belt, [12] North Polar Zone, [13] South South Temperate Zone, [14] South Temperate Zone, [15] South Tropical Zone, [16] Equatorial Zone, [17] North Tropical Zone, [18] North Temperate Zone, [19] North Temperate Zone, [20] the Great Red Spot, together with its associated Hollow. The region between the south edge of the North Equatorial Belt and the north edge of the South Equatorial Belt is System 1, the rest of the planet System 2 has a rotation period that is on average five minutes longer. The belts show marked variations in intensity of colouring and of size, for instance the South Equatorial Belt is sometimes as broad and dark as the North Equatorial Belt.

The clouds on Jupiter are arranged in distinct bands parallel to the equator. The bright "zones" are streamers of ammonia cloud girdling the planet at a higher level than the darker "belts", and this alternating pattern seems to be related to the winds on Jupiter. At the equator, winds blow eastward at a speed of 400 km/h (250 mph), so fast that the equatorial clouds travel once round the planet in only 9 hours 50 minutes – five minutes less than the rotation period at higher latitudes. At higher latitudes too, there is an alternating succession of westerly and easterly winds. The system of wind speeds persisted unchanged during the four months between the two Voyager encounters with Jupiter, even though the cloud patterns had altered quite considerably. The circulation of Jupiter's atmosphere must thus occur not just in the atmosphere where the clouds lie, but must have roots stretching deep within the liquid body of the planet.

The Great Red Spot and other eddies

As well as cloud belts and zones, Jupiter displays enormous numbers of "spots". The largest is the Great Red Spot, which has been observed from the Earth for more than 300 years, and is easily seen with a small telescope. The Great Red Spot varies in size and colour, but its surface area is usually much greater than that of the Earth. Space probe photographs have revealed many smaller rotating spots and eddies, the smallest of them lasting only one or two days, and analysis of the Voyager 1 and 2 photographs has shown that it is the eddies which drive the high-speed winds on Jupiter. Just as a rotating car wheel pushes the car forward and the ground backwards, they drive the gas streams on either side.

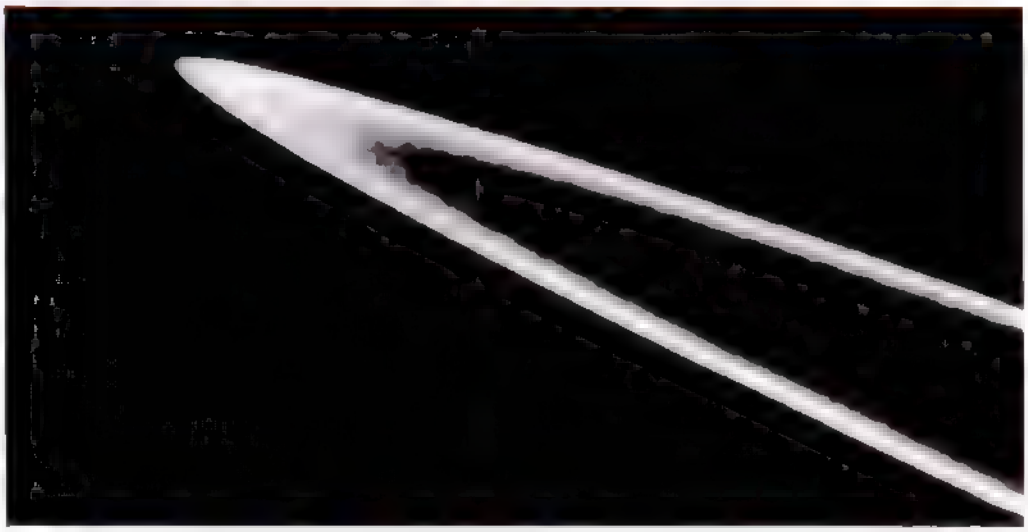
The Great Red Spot is probably just the largest and most persistent of these eddies. Its red colour is however still something of a mystery. The most likely explanation is that the slowly rotating spot is bringing up gases from deep down in Jupiter and that one of the trace constituents is phosphine (a compound of phosphorous and hydrogen). As the phosphine emerges into sunlight, it is broken up and liberates red phosphorous which gives the spot its distinctive colour.

KEY

Jupiter was aptly [illegible] after the king of the Roman gods by the ancient astronomers although they had no idea of the true size of the planet's dimensions or of the number of its satellites. It is larger than any of the other planets [illegible] although it has a diameter about one tenth of that of the Sun. Two of Jupiter's satellites are approximately the same size as the smallest of the principal planets Mercury and it has more satellites (14) than any planet. Shown here are the comparative sizes of the four planets together with [illegible] the Sun and Jupiter. Despite its great distance Jupiter is a prominent object in the sky.

5 Jupiter's Great Red Spot was a prime target for the [illegible] Voyager spacecraft. Early theories had suggested it was a vast [illegible] of stagnant gas above and floating in Jupiter's atmosphere or a huge hurricane. The Pioneer and Voyager probes showed it is a slowly turning anticyclone differing only in size and colour from the shorter-lived white ovals in Jupiter's cloudy atmosphere. The oval has lasted for a few decades, while the Great Red Spot has been [illegible] since the 1660s.

6 Jupiter is circled by a faint ring photographed in detail by Voyager 2 on 10 July 1979 at a distance of 1.5 million km (930,000 miles). It was first detected by Voyager 1 four months before it broke Saturn's much brighter rings of large icy blocks. The ring of Jupiter is composed of very small icy particles. With in the main ring an even fainter ring goes down to Jupiter's cloud tops (off to the right). There are at least two, probably three, moons orbiting Jupiter near the outer edge of the main ring, part of its satellite system.



The moons of Jupiter

Jupiter has at least 16 satellites, which range in size from the largest moon in the Solar System, Ganymede, to small fragments only a few kilometres across. The satellites of which we have some knowledge fall into four equal groups

The four largest moons, and forming one group, are the Galilean satellites, each rivalling or surpassing our Moon in size, with a diameter of 5,276km (3 278 miles) Ganymede is larger than the planet Mercury The Galilean satellites are so bright that they would be visible to the naked eye if they were not overpowered by the glare of Jupiter itself Galileo (1564-1642) saw them when he first pointed a telescope at Jupiter in the winter of 1609-10 The German astronomer Simon Marius (1570-1624) had seen them ten nights before Galileo, but without realizing their nature, and there was considerable dispute over priority. The names now used Io, Europa, Ganymede and Callisto are those suggested by Marius

A small telescope will reveal the changing positions of the Galilean satellites from night to night, and the interesting occasions when they transit across Jupiter's face, or pass behind the planet or into Jupiter's shadow

The smaller satellites

Closer in to Jupiter lies a group of smaller moons. The largest is irregular-shaped Amalthea, which measures 270 by 150km (170 by 90 miles). It was the last moon in the Solar System to be discovered by eye observations at a telescope, by the keen sighted astronomer E E Barnard (1857 1923) in 1892. (Later discoveries have come from telescopic photographs or space probe pictures)

Voyager 1 and 2 photographs revealed an 80-km (50 mile) diameter satellite 1979J2, within Amalthea's orbit, and closer still to the planet, evidence for another two satellites about half this size, designated 1979J1 and 1979J3 The orbits of these last two appear to lie just at the outer edge of a ring of fine dust particles which Voyager 1 discovered encircling Jupiter at a distance of 57,000km (35,000 miles) above the planet's cloud tops

Outside the orbits of the Galilean satellites there are four small satellites in [illegible] but fairly circular, orbits They range in diameter from about 5 to 80km (3 to 50 miles) The latest of these, Leda, was found by Charles Kowal in 1974 he reported another satellite the following year but this has not yet been confirmed

The outermost four moons are also small They travel round Jupiter in the opposite direction from the others, and their orbits are so perturbed by the Sun that they are not even roughly spherical Most astronomers believe that they are captured asteroids or comet nuclei

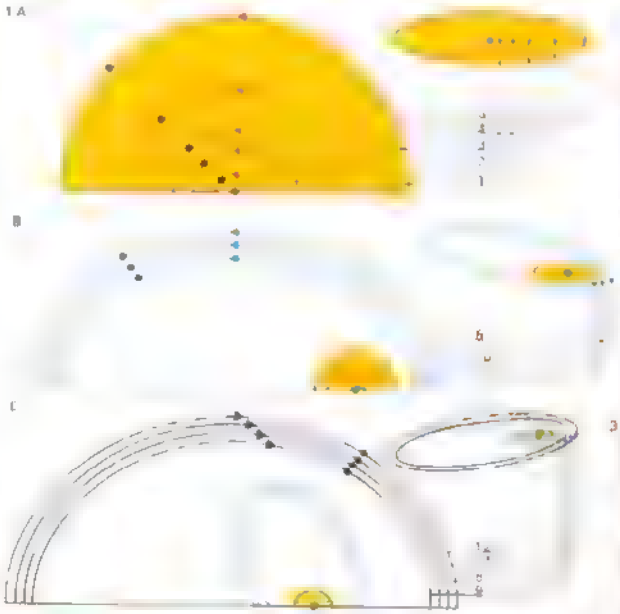
The rock-ice moons

The Voyager 1 and 2 spacecraft investigated the four Galilean satellites in detail as they passed Jupiter in 1979. These moons have turned out not to be dead lumps of rock, but worlds displaying geological activity

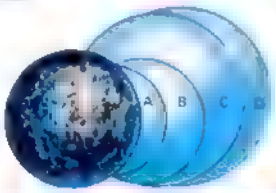
An important clue is given by their density. The Voyager results show that the densities increase from the outer Galilean satellite, Callisto, in to Io Callisto (density

CONNECTIONS

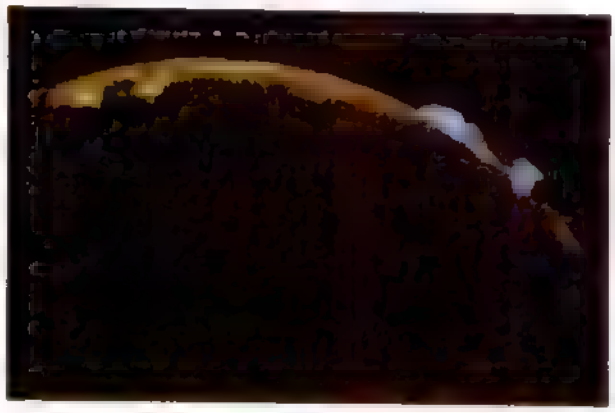
See also



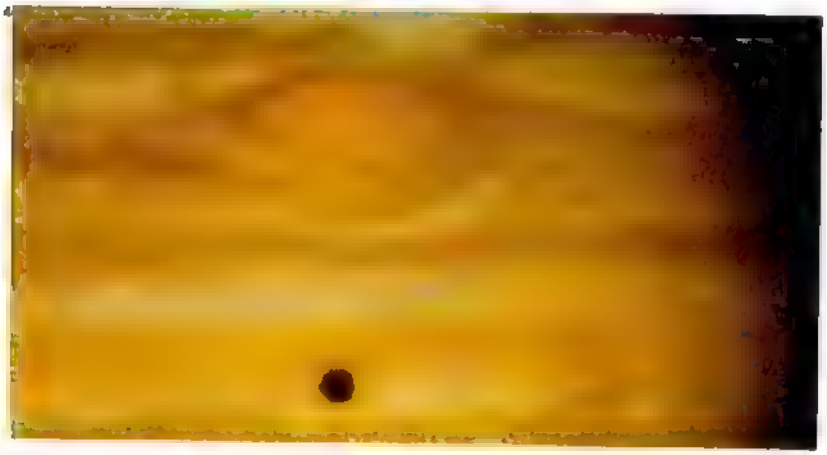
1 The satellites of Jupiter la[illegible] to several well defined groups In the first group (A) the satellite Amalthea (1) seems to [illegible] a mass of [illegible] the diameter [illegible] only about 2[illegible] 124 miles) so that in size [illegible] asteroid Then [illegible] satellites [illegible] discovered by Galileo from 1609 10 – Io (2) Europa (3) Ganymede (4) and Callisto (5) – with mean distances from Jupiter ranging between 422 000km 262 000 miles, for Io out to 1 880 000km 1 170 000 miles) for Callisto The next group (B) consists of [illegible] satellites (6 10 7, plus a 13th and 14th recently discovered the latter unconfirmed) and the third (C) has four retrograde satellites 12 11 8 9)



2 The four Galilean satellites – Europa (A), Io (B), Ganymede (C) and Callisto (D) are compared in size with the moon Europa is smaller but Io is slightly larger than the Moon with a diameter of more than 3 600km (2 23[illegible] miles) Ganymede and Callisto are more nearly comparable with Mercury Ganymede's diameter is 5 000km (3 100 miles) but Mercury's is 4 880km 3 032 miles.



3 The plumes of gas from Io's volcanoes shine blue at the narrow sunlit edge of the satellite seen at [illegible] phase Io is the [illegible] from the Earth known to have active volcanoes and the [illegible] March 18 [illegible] A navigation expert Linda Morabito, was inspecting pictures of the edges of the satellites Io and Europa as an aid to finding the probe's exact position As she turned up the contrast on a photograph [illegible] she noticed [illegible] above the satellite's limb Morabito ruled out the possibility it [illegible] be another [illegible] faint arc lay immediately above a volcano-like [illegible] which had already been spotted on Io's surface She had [illegible] that she had [illegible] the plume of an active volcano Other Voyager 1 photographs [illegible] one had passed when Voyager 1 passed [illegible] month later The [illegible] quarter [illegible] of this gas [illegible] space and forms a [illegible] around [illegible]



4 When a satellite passes in transit across Jupiter it may be seen as a bright spot The two inner large satellites (Io [illegible]) [illegible] easier to see in transit than Ganymede or Callisto because [illegible] higher albedos, or reflecting power Shadow transits are more striking in this photograph taken with the 61in (155cm) reflector at the Catalina Observatory in Texas, the shadow of Ganymede can be seen as a prominent black spot

5 The Galilean satellites' orbits make them easy to observe They may be eclipsed by the planet (1) they may be occulted (2) [illegible] shadows may be seen in transit across the [illegible] Observations of Callisto [illegible] most of the [illegible] are less [illegible] distance from Jupiter

1.81 that of water) and Ganymede (1.93) have too low a density to be made of rock they must be a mixture of roughly equal amounts of rock and ice. Europa's density (3.40) indicates it is mainly rock, with a little ice, while Io, with a density of 3.55, must be entirely rocky. Their differing compositions are probably a result of the copious amounts of heat that Jupiter generated when it formed: the satellites closest to the planet were warmed most, and consequently more of the ice content was evaporated away.

The rock-ice worlds, Callisto and Ganymede, are totally pockmarked with craters, the scars from the impacts of the last of the smaller fragments that built up to form the satellites. The surface of Callisto is the oldest unchanged surface known in the Solar System. Ganymede's crust has since split open to create long parallel grooves between the cratered regions.

Europa is the smoothest body in the Solar System: its icy surface is, to scale, as smooth as a billiard ball. Possibly its ice content melted soon after its formation, and gushed up over its surface to form a world-wide ocean. This has subsequently re-frozen to form a brilliant icy surface.

The volcanoes of Io

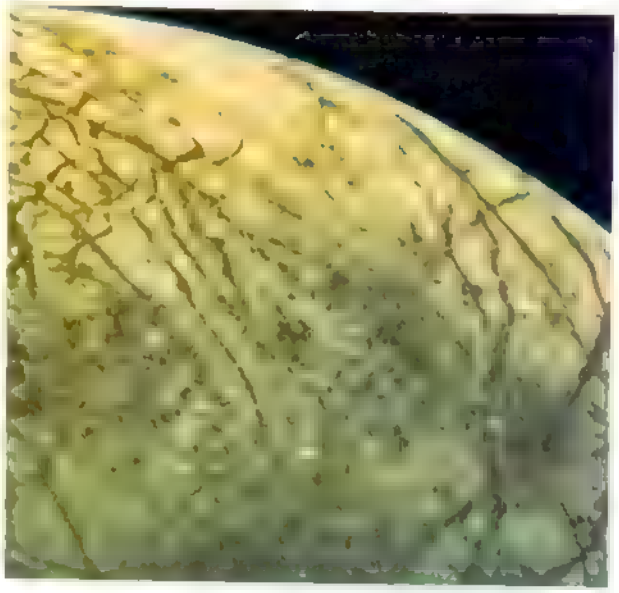
The Voyager 1 pictures of Io showed that this satellite has several active volcanoes. Apart from the Earth, it is the only known volcanically active object in the Solar System.

Io is too small for radioactive heat to melt its rocks. Instead, the responsibility seems to lie with the changing pull of Jupiter together with Europa's gravitational pull on Io, which recurs periodically as the two satellites follow their orbits round Jupiter. This 'stretching' could melt Io's interior.

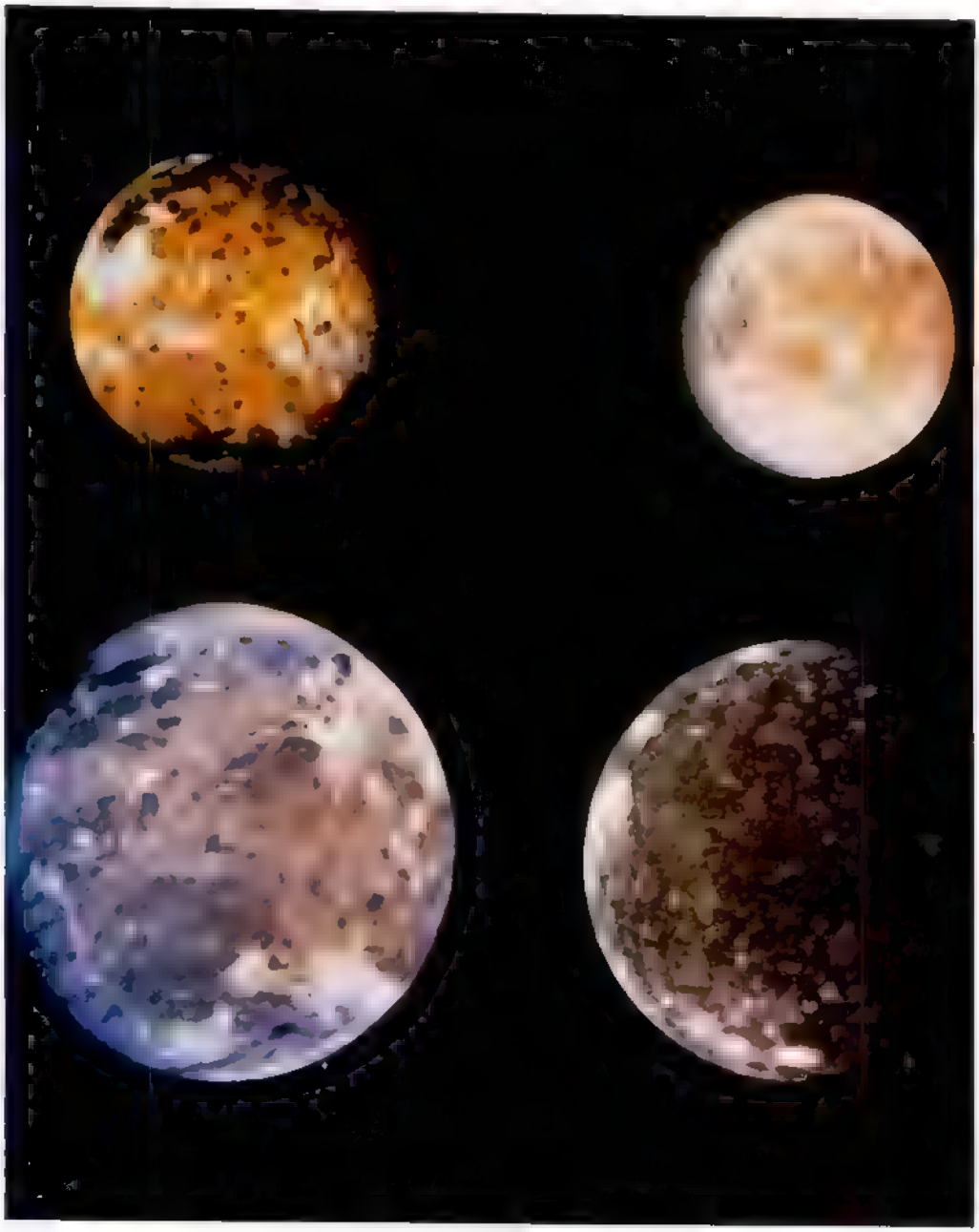
Voyager 1 discovered eight eruptions on Io, each blowing a plume of gases to a height of 70km (45 miles) or more. Volcanic activity seems to come and go, however.

The volcanoes' plumes spread sulphur and sulphurdioxide 'snow' over Io, and the volcanic vents release streams of molten sulphur. Between them, these volcanic products completely resurface Io in less than a million years, and they have completely obliterated the craters that Io once bore.

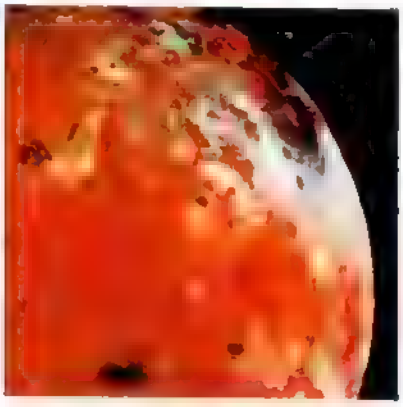
KEY



The bright expanse of Europa's remarkably smooth icy surface is crossed by a network of long dark lines.



6 Jupiter's four large satellites revealed very different characters to the cameras of the two Voyager probes. This composite of Voyager photographs shows them to the same scales of size and brightness. The variation in surface appearance has resulted from their differing geological histories. Dark Callisto (bottom right) has changed little since its birth. Largest moon, Ganymede (bottom left) shows areas of brighter, grooved terrain between the old dark cratered regions. Smallest of the four moons, Europa (top right) has a brilliant icy surface as smooth as a billiard ball. Io's volcanoes have completely coated slightly larger Io (top left) with orange-red sulphur.



7 The surface of Io is totally covered by volcanic debris. The orange sulphur deposits and white regions of solid sulphur dioxide 'snow' have rained down from the plumes of active volcanoes. The darker spots are solidified lakes and streams of sulphur 'lava' which indicate the volcanic vents. The largest of Io's volcanoes (bottom centre) shows as a gently sloping orange cone with dark patches of lava surrounding the peak. Even before the fly-by, some planetologists had suggested that the interior of Io might be melted by the periodic gravitational pull of Jupiter and of the next out satellite, Europa, which travels round Jupiter in exactly twice Io's period, but no one expected erupting volcanoes on its surface. Volcanic outbursts are building up thick layers of new surface on Io, so that no part we see is more than a million years old, and as a result no impact craters are visible. It may well have a crust of frozen sulphur lying over an ocean of liquid sulphur: truly an extraordinary world.

The giant planet Jupiter and one of its satellites, photographed from a range of [illegible] km [illegible] miles) by the Voyager probe. Jupiter's great Red Spot is visible among the complex cloud patterns of its southern hemisphere. The orange [illegible] rotated from [illegible] visible in front.

This montage of Voyager 1 photographs not to scale shows Saturn and some of its 17 moons as the foreground is heavily cratered 132km [illegible] diameter Dione above are Tethys and Mimas with its huge crater to the left is Enceladus

The planet Saturn

Saturn is the most distant planet easily visible to the naked eye, and it is a giant world second only to Jupiter in size and mass. At its distance of 1,427 million kilometres (887 million miles) from the Sun Saturn takes 29.46 years to complete one orbit. In our skies, Saturn appears as a dull-yellowish "star", creeping very slowly through the background constellations. A moderate telescope will however reveal Saturn's globe and beautiful rings.

Saturn itself is 120,000 km (74,500 miles) in diameter – 9.5 times the diameter of Earth – and like Jupiter, it is flattened by its rapid rotation. Saturn turns once in 10hr 39min, although clouds go rounds its equator in only 10hr 14min because of the strong winds there. The planet is less than one-third as massive as Jupiter, but at 95 Earth masses Saturn still outweighs all the remaining planets put together. Its size and mass mean that its density is less than that of water: given a large enough ocean, Saturn would float in it!

Like Jupiter, Saturn is made up mainly of hydrogen and helium, and it has a similar internal structure: a small, rocky core, a surrounding layer of liquid metallic hydrogen that generates a magnetic field 1,000 times stronger than the Earth's, a larger region of liquid hydrogen-helium mixture, and at the outside, a relatively thin atmosphere containing clouds.

The rings as seen from Earth

Saturn's rings are one of the most beautiful sights to be revealed by an amateur astronomer's telescope. They stretch 272,000 km (170,000 miles) from edge to edge. A moderate telescope will show that Saturn has two main rings separated by a dark gap, the Cassini Division, named in honour of its discoverer Giovanni Cassini (1625-1712). The A-ring lies outside the Cassini Division, and the B-ring within. Inside the B-ring is the fainter C-ring.

Space probe techniques have indicated a still fainter D-ring within the C-ring, and a broad but faint E-ring well outside the A-ring. Space probes have shown that Saturn has two very narrow rings lying between the A- and E-rings, both too narrow to be seen from Earth. Pioneer 11 found the F ring in 1979, the Voyager 1 the G ring in 1980.

The ring system lies in the plane of Saturn's equator, and because this is tilted at an angle of 29 degrees, the appearance of Saturn as seen from the Earth changes continuously as Saturn moves around the Sun. At times when Saturn's ring system is presented edge-on, as it was in 1966 and 1979-80, it cannot be easily seen with even the largest telescopes on Earth. This indicates that the rings are less than a kilometre (0.5 mile) thick. Spectroscopic observations from Earth also show that the rings are not continuous sheets of matter, but must consist of millions upon millions of tiny particles orbiting independently as "micro-moons".

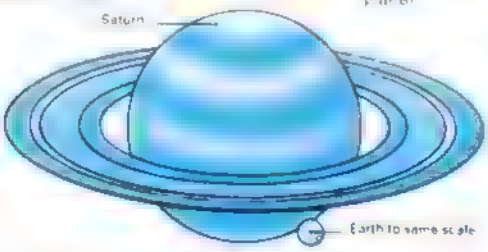
The Voyagers' view of Saturn

Further information had to await space probe missions. The trail-blazing Pioneer 11 passed Saturn in 1979, discovering its magnetic field and the F-ring, but the real breakthrough came with Voyagers 1 and 2 in 1980 and 1981. They sent back more than 30,000 high-resolution photographs

CONNECTIONS

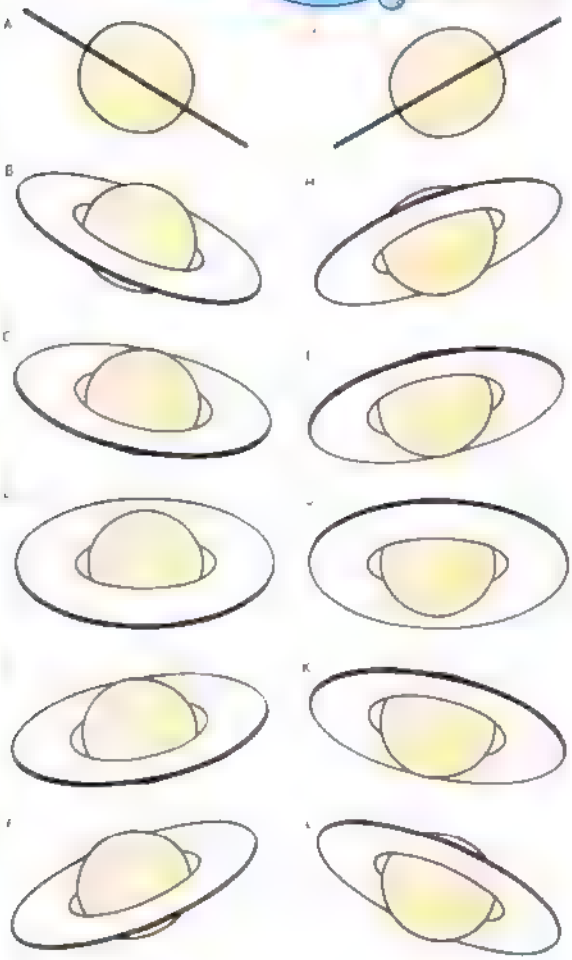
See also

1 Saturn is a giant planet, although not as large as Jupiter, its volume is 744 times that of Earth. In this scale diagram the Earth would fit neatly into the gap between the Crêpe Ring and the surface of the planet.

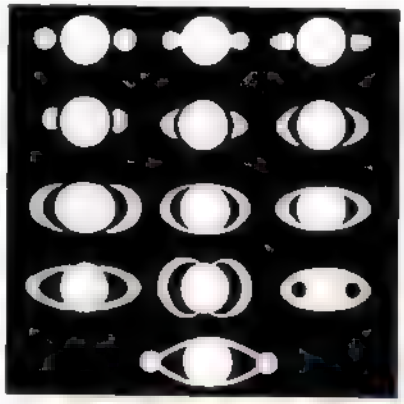


2 The mean density of Saturn is [illegible] only [illegible] that of water. This is much less than that of any other principal planet. It has been [illegible] that if Saturn were dropped [illegible] it would float. The low density is due to the [illegible] of the [illegible] elements, hydrogen and helium. The [illegible] show the [illegible] of the Earth and Saturn [illegible] compared with the density of water.

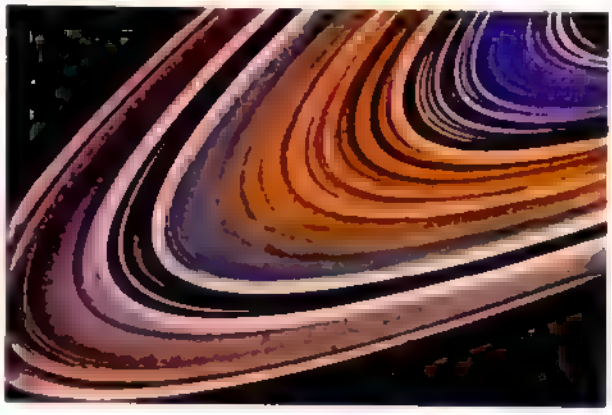
3 The inclination of Saturn's axis to the plane of the [illegible] 26°44' [illegible] greater than that of the Earth. The rings [illegible] the planet exactly in the plane of the planet's [illegible]



4 The aspects of Saturn's rings vary considerably as seen from Earth. At regular intervals as the rings [illegible] the plane of the Sun and Earth (A). The rings then seem to [illegible] (B) they are [illegible] maximum advantage (D) after which they close up again. When the south pole of Saturn [illegible] towards the Sun, the southern ring face is displayed and at such times part of the northern hemisphere of the planet [illegible] these diagrams [illegible] Subsequently it is the northern [illegible] which [illegible] to observers [illegible]



5 To early telescopic observers Saturn was an observational enigma. The first recorded drawings seem to have been made by Galileo (1564-1642), whose feeble telescope was [illegible] to show the ring system in its true guise. He believed [illegible] altogether because during that time they had [illegible] to the Earth.



6 The colours of Saturn's rings have been exaggerated in this photograph taken by [illegible] probe Voyager 2 so that various shades show subtle differences in the [illegible] composition. Many scientists [illegible] least [illegible] show [illegible] the [illegible] A and B rings which [illegible] shaded [illegible] to [illegible] has [illegible] others [illegible]

Saturn appears a much blander world than Jupiter because its weather systems are obscured by a thick overlying layer of haze. But computer-enhanced Voyager photographs reveal a rather similar pattern beneath the haze: coloured bands at different latitudes, rotating spots and the heads of convective clouds. The main difference is that Saturn's equatorial winds are even faster than Jupiter's – 1,800 km/h (1,100 mph), which is about two-thirds the speed of sound on Saturn. These winds quickly break up the cloud patterns, so there are no prominent spots near the equator.

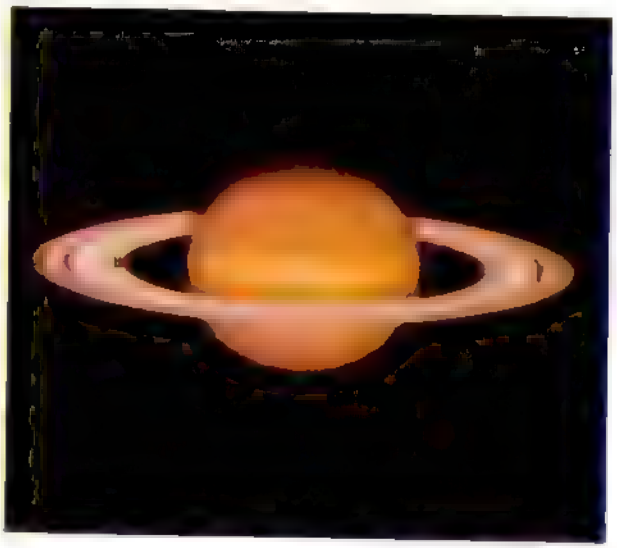
As with Jupiter, Saturn's weather seems to be driven by a combination of the Sun's heat and extra heat produced within the planet itself. Unlike Jupiter, though, Saturn can hardly have retained any of the original heat produced during its formation. The extra energy may instead come from the separation of hydrogen and helium in the liquid layers within Saturn's globe.

The Voyagers' main surprises – and their most spectacular photographs – came from their studies of Saturn's rings. It turns out that the main A-, B- and C-rings are not smooth expanses of micro-moons, but are broken up into hundreds of thousands of narrow "ringlets". Each ringlet is less than a kilometre (0.5 mile) wide, and is separated from its neighbours by an equally narrow gap. In other words, the micro-moons constituting the ring system cannot orbit Saturn at any arbitrary distance, but must follow nose-to-tail in orbits at certain particular distances from the planet.

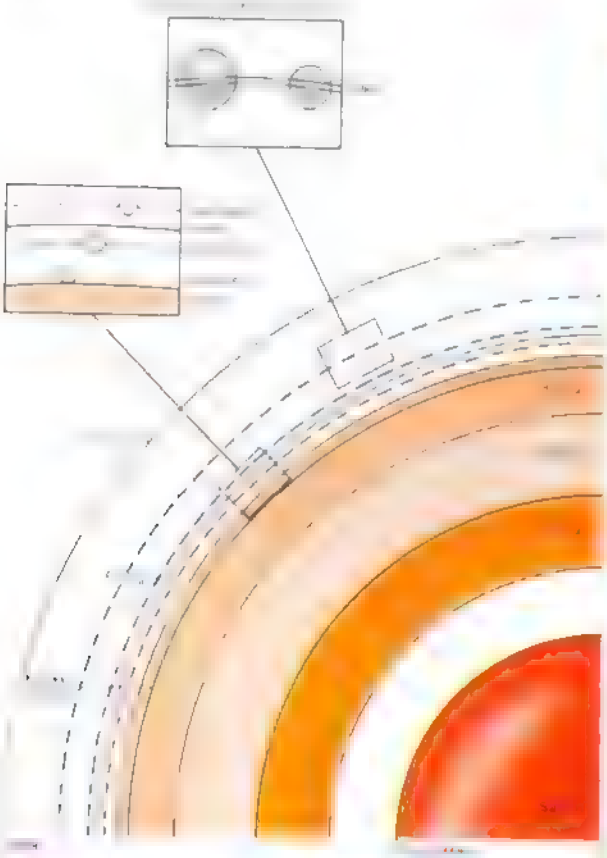
The origin of the ringlet structure is still unexplained. It may be due to little-understood gravitational interactions between all the same micro-moons.

Analysis of radio signals from Voyager 1 shows that the average size of the micro-moons in the C-ring is about 2m (6ft), and in the A ring 10m (33ft).

Another major surprise was Voyager 1's discovery of "spokes" – dark bands lying across the rings. Their light scattering properties show that they are made up of tiny dust or ice particles. Their origin and behaviour are not fully understood but Saturn's magnetic field may well be responsible.



The southern hemisphere of Saturn is displayed in this photograph. Pa



7 The complexity of Saturn's ring system has only become evident with the [illegible] from the Voyager [illegible] The three main rings A, B and C have long been known from earth observations. A and B are separated by the Cassini division [illegible] thought to be empty space. The B-ring [illegible] now turns out to contain several [illegible] faint ringlets. The B-[illegible] particles [illegible] that the C-ring [illegible] material here [illegible]

[illegible] seen from Earth. The [illegible] System S15 (1980 [illegible] S1 [illegible] S2 and S3. S2 [illegible] shepherd [illegible] Voyager

8 This parting shot taken by [illegible] of 7 November [illegible] B and C rings.

9 Colour-enhanced view [illegible]

The moons of Saturn

Saturn [illegible] 11 known satellites, vals [illegible] of Jupiter, and is in some ways equally interesting.

The largest moon, Titan, can be seen with a [illegible] telescope. It was discovered in [illegible] by the Dutch astronomer Christiaan Huygens (1629-1695). Measurements from the Voyager I space probe show that Titan is 5,120km (3,180 miles) across, so it is second only in size to Jupiter's Ganymede.

Giovanni Cassini (1625-1712) found the next largest satellites: 1,440km (900 miles) diameter Iapetus and 1,530km (950 miles) Rhea, in 1671 and 1672 respectively, and in 1684 two satellites closer in to the planet, 1,120km (690 miles) diameter Dione and Tethys [illegible] 1,050km (650 miles) across. Just over a century later, the great astronomer William Herschel (1738-1822) detected Mimas and Enceladus, which are smaller in size and smaller, with diameters of 300km (240 miles) and 500km (310 miles) respectively. The nineteenth century saw the discovery of the outer moons. Hyperion, roughly the same size as Mimas, and tiny Phoebe, only 200km (120 miles) across, [illegible]

French astronomer Audouin Dollfus (1924-) reported the discovery of another small satellite even closer to Saturn than Mimas when the rings appeared edge-on in 1966. He called this moon Janus. Later analysis, however, has shown that there is no satellite in this orbit, and the name Janus has been dropped. There are, however, two small satellites, 1980 S1 and 1980 S3, which move in almost the same orbit as each other.

The Voyagers resolve the smaller moons

As the Voyager space probes passed Saturn in 1980 and 1981, they were able to check out suspected satellites and discover new ones.

The closest in satellite of Saturn, 1980 S28, orbits the planet only 800km (500 miles) outside the edge of the A-ring. The gravity of this 60km (40-mile) wide moon keeps the outer edge of the ring system in check. The next two, the 'shepherd moons', may well influence the shape of the F-ring.

Farther out lie the co-orbital satellites. These two irregular moons, with long axes of 220km (140 miles) and 140km (90 miles), follow orbits at almost exactly the same distance from Saturn and periodically seem to exchange orbits. They are almost certainly the halves of a moon which was broken in two long ago.

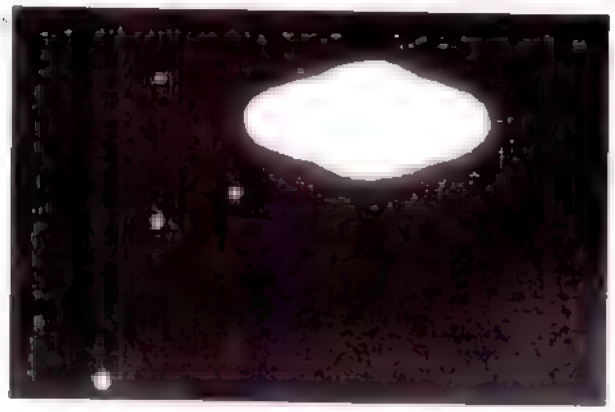
Voyager 2 confirmed suspicions from Earth-based observations that tiny satellites, only 60km (40 miles) in size, lie in the same orbits as Tethys and Dione. Dione has one such "trojan" moon lying in front it, one sixth the way round its orbit. Tethys has two "trojans", one ahead and the other an equal distance behind. In these positions, a small satellite is perfectly stable, like the Trojan asteroids that share Jupiter's orbit.

Ice-ball worlds

Between them, the two Voyager probes also investigated all the previously-known satellites of Saturn. Except for Titan, all the satellites whose density could be measured turned out to have between 1.0 and 1.4 times the density of water. They must consist largely of ice, with only a limited admixture of rocky material.

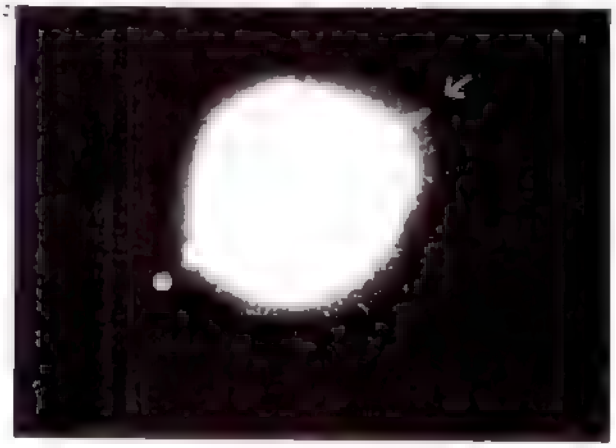
CONNECTIONS

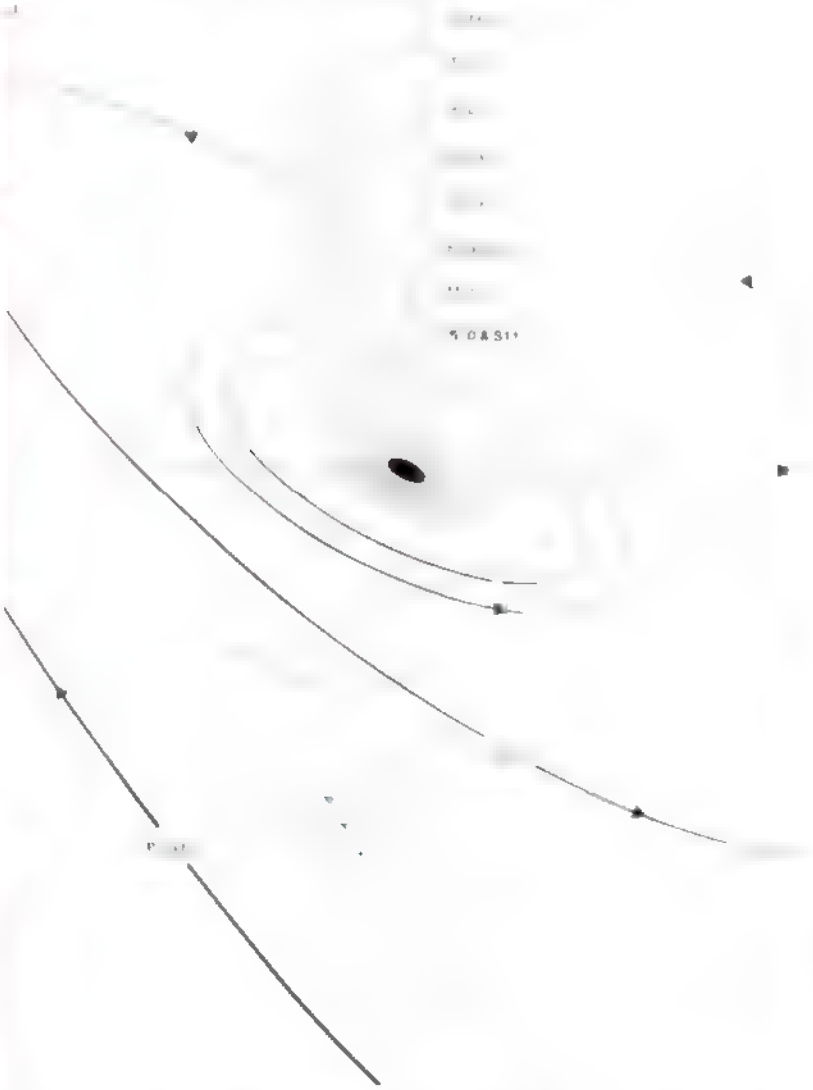
See also



1 The inner satellites of Saturn [illegible]

2 Earth based pictures [illegible]





3 The orbits of Saturn's satellites [illegible] plane of the planet's equator, which is also the plane of the ring system [illegible] Phoebe is shown lower left with Hyperion and Iapetus. Phoebe has retrograde motion and may be a captured asteroid. In 1905 William Pickering (1858-1938) reported another satellite between Titan and Hyperion, but it has not been seen since. He possibly mistook a star for a satellite.

Most of them are heavily cratered by impacts dating from their birth. Mimas and Tethys are remarkable in displaying craters fully one third as large as the satellites themselves. The impacts that created these craters must have almost split the moons in two. Parts of the second-largest moon Rhea must have been covered with new surface material soon after its birth, but oddly enough, the much smaller moon Enceladus has been partially resurfaced much more recently, and mild geological activity may still be continuing there.

The outermost moons each have individual quirks. Although Hyperion is the twin of Mimas in size, it is not spherical but "hamburger-shaped" 410 by 220km (250 by 140 miles). On the other hand, the outermost moon Phoebe is small – 200km (120 miles) across – but it is spherical. Phoebe is also very dark in colour, and it follows a very large retrograde orbit. It is probably an object captured as it passed Saturn, an asteroid or huge comet nucleus.

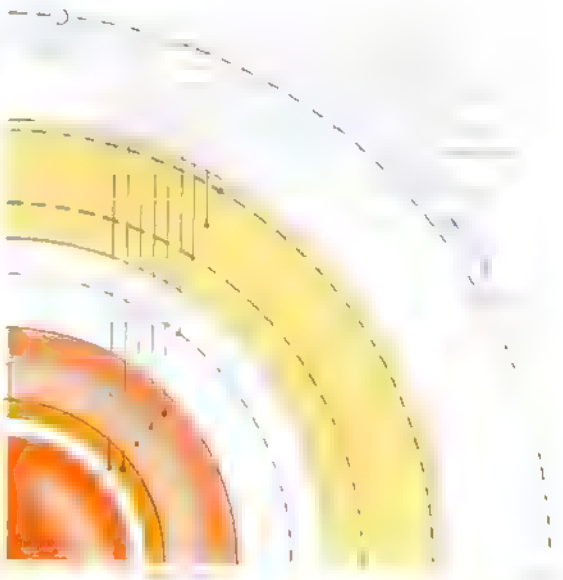
Iapetus is a bright ice ball world with one hemisphere coated with a black deposit as dark as pitch. This may be dark matter shed from Phoebe, more probably it welled up from inside Iapetus itself.

Orange-clouded Titan

Earth-based observations in the 1940s showed that Saturn's largest moon, Titan, has an atmosphere – the only satellite known to posses one.

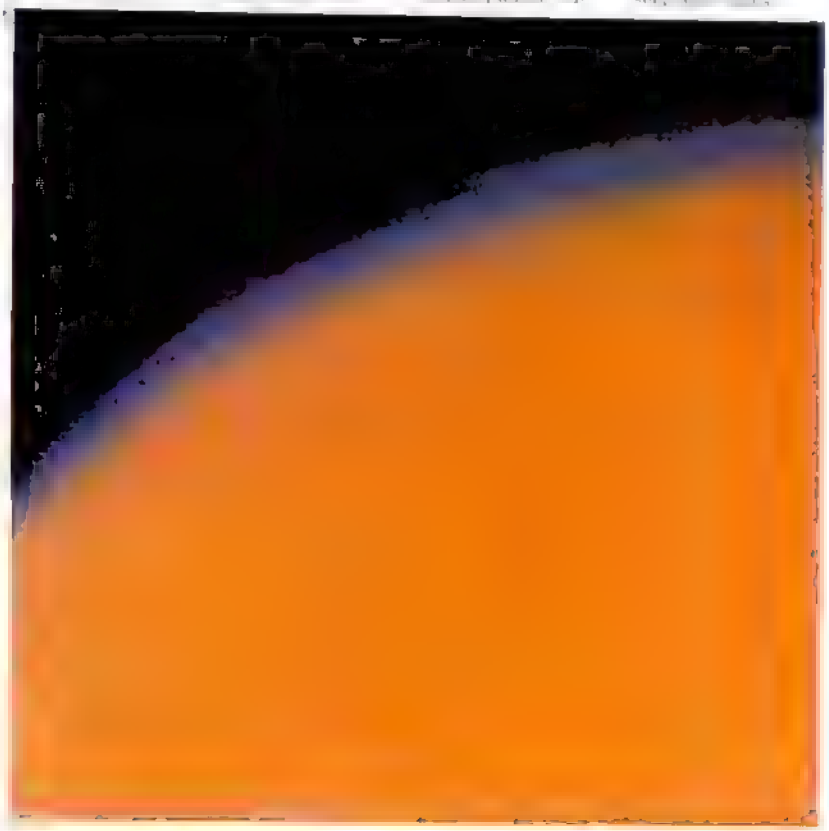
Voyager 1 discovered that the atmosphere consists mainly of nitrogen, with smaller amounts of methane, ethane, acetylene, ethylene and hydrogen cyanide. The atmosphere is dense and pressure high.

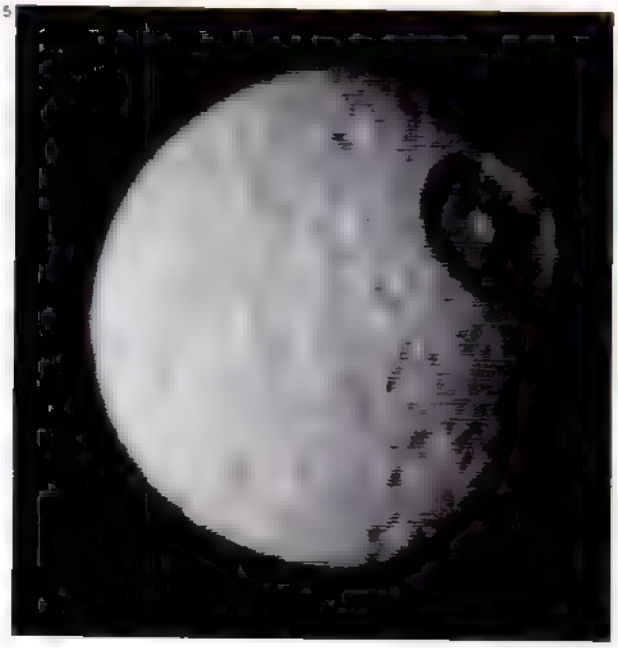
Unfortunately, the surface of Titan is not visible. The Voyager photographs show that the entire world is enveloped in a thick layer of orange haze, probably composed of complex organic molecules formed by reactions between the simple organic molecules in the atmosphere. Indirect measurements, however, indicate the surface must be at a temperature of -180°C (-292°F). In these conditions, methane is near its solidification point. Titan may well have methane rainfall, methane rivers and oceans.



Two of Saturn's major satellites orbit within the ter[illegible] F ring. Dione [illegible]

4 The first view of Saturn's F-ring from the Voyager [illegible] came as a shock to astronomers. Taken on 12 November 1980 at a range of [illegible] 450,000 miles, the picture showed that the F ring consists of three narrow ringlets and that two of these are [illegible] orbit Saturn just inside and [illegible] orbit Saturn, just inside and [illegible]





5 The innermost major moon Mimas [illegible] 390km [illegible] diameter and [illegible] density shows that it [illegible] Mimas shows a crater more than [illegible] miles across, almost one third the size of Mimas itself. The [illegible] has the haze [illegible] which has suffered this late [illegible]

6 A close-up picture [illegible] haze (blue, stretch[illegible] [illegible]

The outer planets

[illegible]

Uranus and its strange tilt

[illegible]

The discovery of Neptune

[illegible]

who became Astronomer Royal at Greenwich the following year, but he showed little interest. John Couch Adams (1819–92), working at Cambridge in 1843, resolved to tackle the problem. He thought that using the perturbations of Uranus it would be possible to find the position of the unknown planet, and after several months of hard work he felt that he had fixed the position accurately. He then contacted Airy, but again the Astronomer Royal took no action, although to be fair this was partly due to a series of misunderstandings. In the meantime similar calculations made by the French mathematician Urbain Leverrier (1811–77) were sent to the Berlin Observatory, where Galle and d'Arrest quickly located the planet in almost the same position.

Neptune is almost twin to Uranus. It is slightly smaller with a diameter of 49 500 km (30 940 miles) and more massive. It does not have Uranus's remarkable axial tilt. The orbital period is 164.8 years and the mean distance from the Sun is 4 497 million kilometres (2 794 million miles). Neptune's telescope will show definite surface detail on

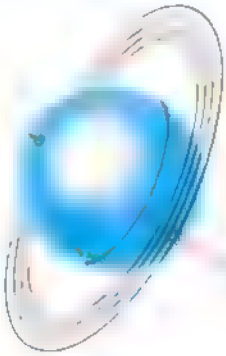
CONNECTIONS

See also

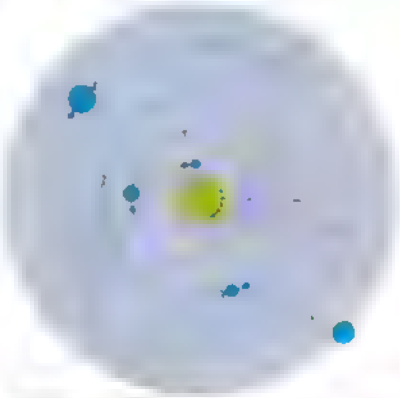
[illegible]

1 Uranus has a den sity [illegible]

2 The rotation period of Uranus [illegible]



3 Orbits of the five satellites of Uranus



4 Uranus was photographed [illegible]

5 Uranus and Neptune [illegible]

6 When Galle and d'Arrest [illegible]

Neptune all that can be seen is a pale bluish disc. Of the two satellites [8] Triton is larger than our Moon and has a circular orbit but moves round Neptune in a retrograde direction; the other satellite, Nereid, is small and has an eccentric orbit.

The enigma of Pluto

Even with the discovery of Neptune there were still discrepancies in the movements of the outer giants, and Percival Lowell (1855–1916) undertook new calculations with the aim of finding yet another planet. In 1930 at the Lowell Observatory in Arizona the planet was duly located [9], although by then Lowell had been dead for 14 years.

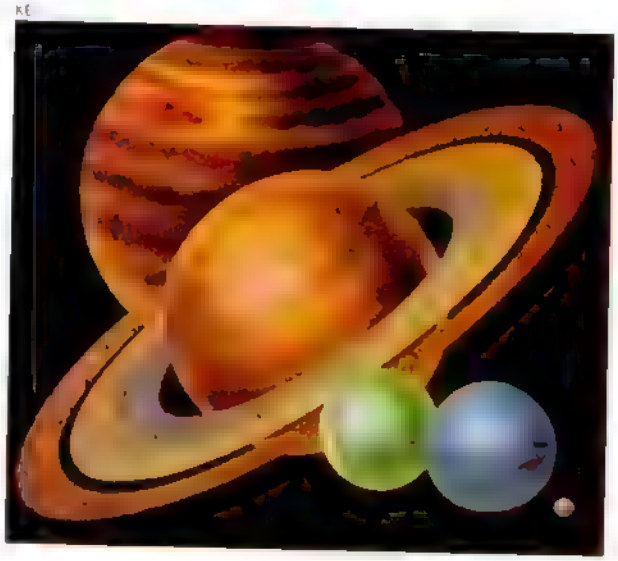
The planet, named Pluto, has the most eccentric orbit of any planet in the Solar System [10]. For twenty years of its 248-year orbit Pluto is closer to the Sun than Neptune; this situation occurs between 1979 and 1999.

Pluto is the smallest planet. It is too small to show a disc in a telescope in normal circumstances, but the special technique of speckle interferometry has recently shown that its diameter is about 3,500km (2,200 miles) – very similar to our Moon. Pluto rotates in 6.4 Earth days.

In 1978, James W. Christy discovered that Pluto has a satellite, Charon, which orbits Pluto at a distance of 20,000km (12,000 miles). Speckle interferometry measurements show that Charon is almost half the size of Pluto itself, so the Pluto–Charon system is more realistically regarded as a double planet or even a double asteroid.

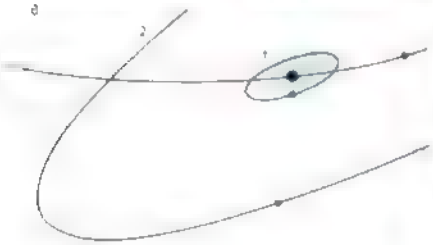
With the discovery of Charon, astronomers could calculate Pluto's mass, which turns out to be only 1/500 the Earth's mass. Pluto's density works out as about the same as water. This is consistent with a planet consisting mainly of frozen methane.

Its low mass means that Pluto could not have caused the perturbations of Uranus and Neptune on which Lowell based his calculations. The discovery was due mainly to the thorough search made by Clyde Tombaugh at Lowell's observatory and only by coincidence was it near the predicted position. The planet which Lowell had predicted may still await discovery.

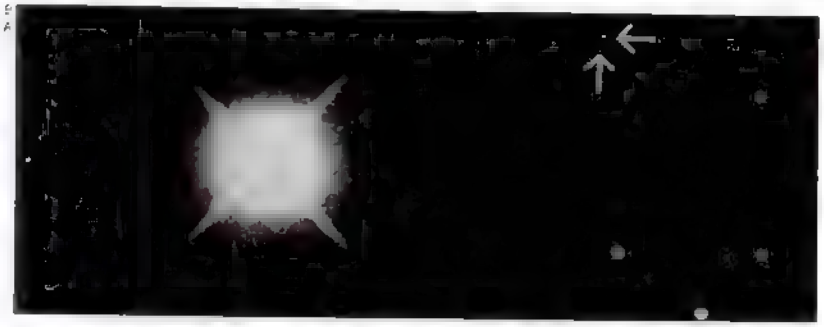


The sizes of the outermost planets are compared here with Jupiter and Saturn. Uranus and Neptune are giants with gaseous surfaces. Pluto is smaller than the Earth and is looked on as a terrestrial planet, although its nature is uncertain.

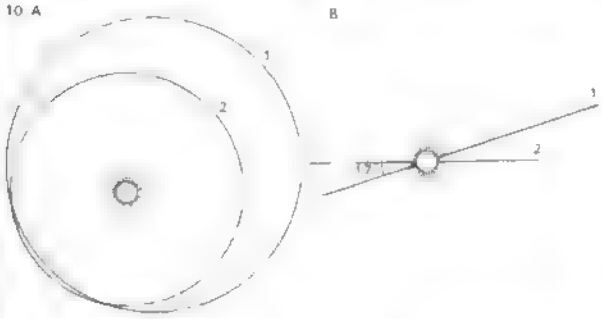
7 This photograph shows Neptune [illegible] Triton near the [illegible] of the satellite's [illegible] of the same [illegible]. Uranus and was discovered by the English astronomer William Lassell a few weeks after the discovery of Neptune itself. Nereid is very faint and can be photographed only with giant telescopes.



8 Orbits of Triton [1] and Nereid [2]. Nereid [illegible] only a large satellite [illegible] Solar System [illegible] in this way. Nereid has a [illegible] orbit [illegible] that of a comet [illegible] grade [illegible] the [illegible]



9 These photographs show the discovery of Pluto in 1930 by Clyde Tombaugh at the Lowell Observatory [illegible] Percival Lowell's calculations. A [illegible] March [illegible] the shift of [illegible] by the grid was [illegible] the exposed image [illegible] that of the [illegible] star Delta [illegible] Pluto's [illegible] of about magnitude 14 and may thus be seen with a telescope of moderate size.



10 Pluto has an exceptional orbit [A] that is both highly [illegible] and highly eccentric [illegible] compared with that of Neptune. 2. At perihelion Pluto may come within Neptune's path but the inclination of 17° means that there can be no collision. The next perihelion passage is due in 1989. Its peculiar orbit has given rise to some doubt as to whether Pluto should be considered a true planet.

Comets

A great comet, with a brilliant head and a tail stretching half-way across the sky, is a spectacular object – and it is easy to understand why comets of this kind caused such terror in ancient times. Comets have always been regarded as unlucky and fear of them is still not dead in some primitive societies.

Yet a comet is not nearly as important as it may look. It is made up of small particles (mainly icy in nature) and tenuous gas. On several occasions the Earth has been known to pass through a comet's tail without suffering the slightest damage.

A comet's anatomy

A large comet is made up of three principal parts: a nucleus (containing most of the mass); a coma or head, and a tail [2]. The coma and tail appear only when the comet approaches the Sun and solar radiation vaporizes the icy nucleus. As the comet recedes the tail disappears. Small comets, however, are often devoid of tails and look rather like small patches of faintly luminous cotton wool in the sky.

The tails of comets are of two main kinds: gas and dust. Generally the gas tail is relatively straight, while the dust tail is curved – because it lags behind as the comet moves forwards. One remarkable feature of tails is that they always point more or less away from the Sun, so that when a comet is receding it travels tail-first. The causes of this are not known for certain, but it is thought that the tiny particles in the tail are repelled by the so-called "solar wind", a stream of electrified, low-energy particles constantly flowing outward from the Sun in all directions.

Comets are members of the Solar System but their paths, in most cases, differ from those of the planets in that they are much more eccentric. Dozens of comets with short-period orbits are known, the period of Encke's comet, for instance, is only 3.3 years, so that it is seen regularly and has been observed at more than 50 returns to perihelion (the point in its orbit when it is closest to the Sun – aphelion is the point farthest from the Sun) from when it was first sighted in the eighteenth century. Comets are not visible through the emission of light of their own, but through the reflection of the Sun's radiation, which causes the cometary material to fluoresce. Thus most comets cannot be followed throughout their orbits and can be seen only when relatively close to the Earth and the Sun.

Short- and long-period comets

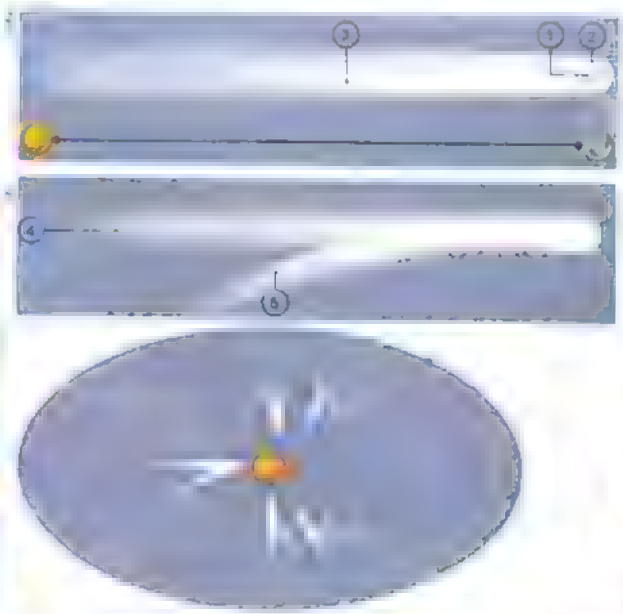
All the short-period comets are faint and many of them are difficult objects to view telescopically. A few (notably Schwassmann-Wachmann 1 and the more recent Gunn's comet) have more circular paths [3] and can be followed throughout their orbits.

Other comets take decades to travel once round the Sun. The most famous of these is Halley's comet [5], which is a bright naked-eye object and is seen every 76 years or so (the period of a comet is not constant because of the perturbations caused by the planets). Halley's comet last returned to perihelion in 1910 and is due back once more in 1986. Records of it go back well before the time of Christ.

Other great comets have much longer periods – so long, indeed, that we cannot measure them accurately. Thus comets of this

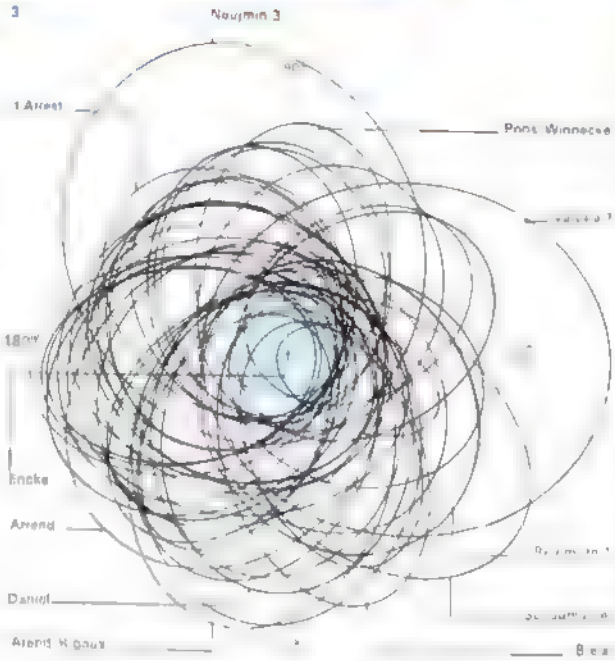
CONNECTIONS

See also



2 The anatomy of a large comet (A) contains the nucleus (1), which may or may not be a conglomerate, the coma (2) made up of small particles and tenuous gas, and the tail (3) extending away from the coma. The two types of tail shown in B are a gaseous tail (4), which is generally straight, and a dust tail (5), which lags behind the moving comet, so that the tail appears curved. The tail of a comet like Halley's (C), always points approximately away from the Sun whatever its orbital position. A comet develops a tail only as it nears perihelion, losing it as it recedes from the Sun.

1 There are three main classes of comet. Short-period comets (A) often have their aphelia at approximately the distance of Jupiter's orbit (1). Their periods amount to a few years and all short-period comets are faint. Long-period comets (B) have aphelia near or beyond Neptune's orbit (2) – Halley's is the only conspicuous member of the class. Comets with very long periods (C) have such great orbital eccentricity that the paths are almost parabolic. Because only a short arc can be measured it is impossible to calculate the periods of these comets accurately. All the really brilliant comets apart from Halley's, are of this type.



3 Some short-period comets, whose aphelia lie near Jupiter's orbit, are said to belong to Jupiter's family. Their orbits are shown in relation to those of Earth (the inner blue ring) and Jupiter (the outer). Encke's comet has the shortest period (3.3 years), that of Schaumasse is more than 8 years. It was thought that comets came from interstellar space and were captured by the planets, but this theory is no longer accepted as valid.

4 Donati's comet of 1858 was generally thought to be the most beautiful ever observed. It was a brilliant naked-eye object and had tails of both gas and dust. This picture is taken from an old woodcut.

kind cannot be predicted and are always apt to take astronomers by surprise. Such was the great comet of 1843, whose coma was larger than the Sun, even though its mass was, by astronomical standards, negligible. Other great comets appeared in 1811, 1882 and 1910. The Daylight comet of 1910 (not to be confused with Halley's) was probably the brightest to have been seen during the present century. Kohoutek's comet of 1973 was expected to be brilliant but proved, in the event, a great disappointment [9].

Short-lived comets

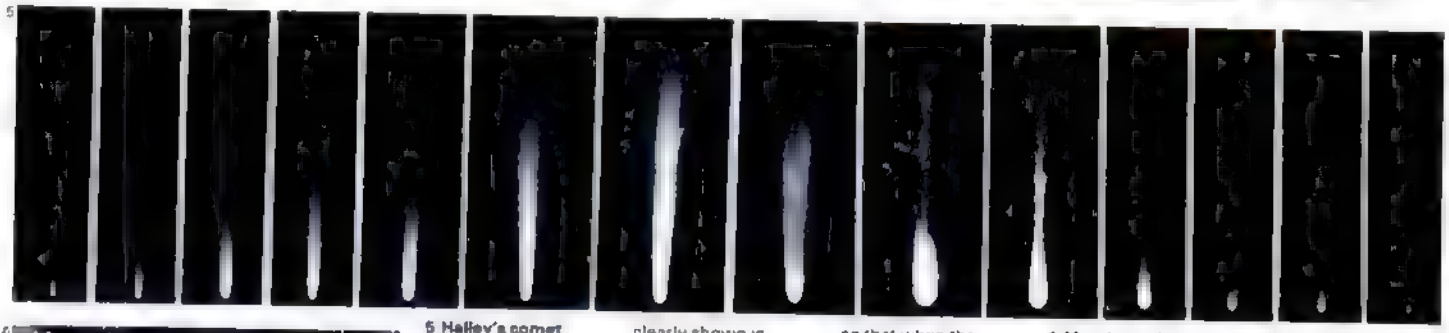
Some great comets approach the Sun very closely and are termed "Sun-grazers". As the comet passes perihelion the tail swings round, and it often happens that the original tail is destroyed and a new one forms. The tails of comets are produced from evaporation from the nucleus and there must be a steady wastage of material, so that by cosmic standards comets are short-lived. We even know of comets that have disappeared. Westphal's comet of 1913, which had a period of 62 years, faded out as it approached perihelion and was never seen again. Biela's periodical comet, which took 6.75 years to complete one orbit, split in two in 1846, the "twins" were seen again in 1852 but that was their last appearance as comets. In 1872, when due to return, a bright meteor shower was seen in the region from which they ought to have come. This emphasizes the close connection between comets and meteors. Meteors may in fact be authentically regarded as cometary debris.

There is considerable uncertainty about the origin of comets. But according to J. H. Oort, a Dutch astronomer, there is a vast "comet cloud" at a great distance from the Sun; sometimes a comet will swing inwards towards the Sun, when it can be observed.

Although professional astronomers are engaged in comet-hunting, many discoveries are made by amateurs. G. E. D. Alcock, an English schoolmaster, has now discovered four, while the bright comet of 1970 was found by a South African, J. Bennett.

Comets are of great scientific interest and there is serious talk of sending a space probe to rendezvous with a suitable comet.

This brilliant comet is Bennett's comet, photographed on 12 March 1970. The tail is quite long and its fine gaseous nature is clearly seen. The nucleus is not shown because of the over exposure of the coma, or head, in order to bring out the structure of the tail. The coma is made up of material from the nucleus which is vaporized by solar radiation as the comet approaches the Sun, when its temperature may rise to several thousand degrees. A comet's emission spectrum reveals the presence of such elements as iron, calcium, sodium, potassium, copper, chromium, nickel and traces of several other metals.



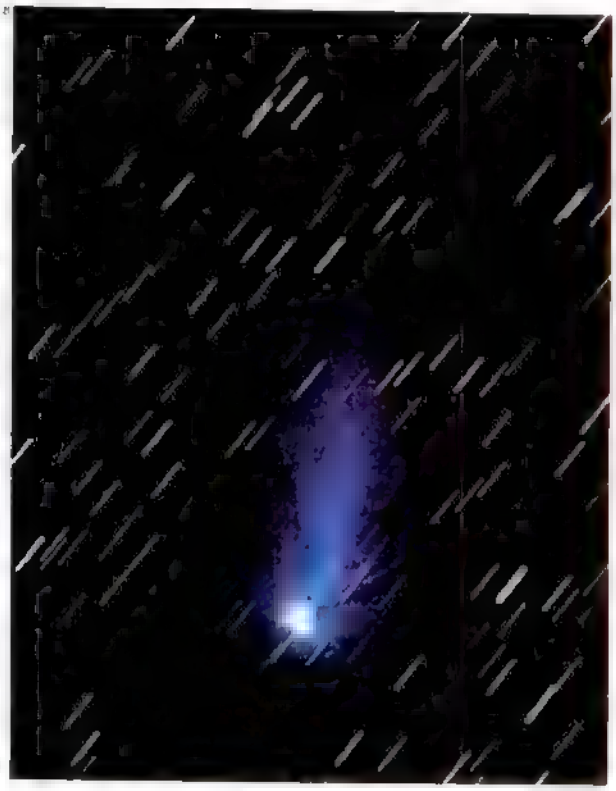
5 Halley's comet last returned to perihelion in 1910 although not as bright as the great "non periodical" comets, the increase and decline of the tail is clearly shown in this sequence. As it approached perihelion the tail developed enormously after the closest approach to the Sun the tail contracted so that when the comet was last seen the tail had disappeared altogether. The seventh picture shows the tail shortly before perihelion passage.

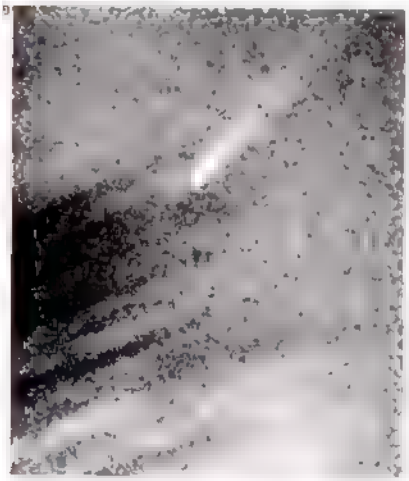
6 Morehouse's comet of 1908 had a complex tail, the structure of which changed rapidly. Great disturbances must have been taking place, but the comet was not bright enough for the details of these changes to be seen on Earth.

7 Comet Arend-Roland of 1957 was one of the most interesting comets of recent times. The apparent "forward spike" is not an extra tail, but is illuminated meteoritic debris lying along the comet's orbit.

8 Humason's comet of 1961, shown in this photograph taken with the 48in (121cm) Schmidt telescope at Palomar, USA, was one of the first comets to be photographed in colour. Because the telescope was tracking the comet surrounding star images appear as short trails.

9 Kohoutek's comet of 1973 was not as spectacular as was hoped. There will be no opportunity to see it again since it will not return to perihelion for 75,000 years.





Meteors and meteorites

Meteors or shooting stars can usually be seen on clear August nights in the Northern Hemisphere. They are rapidly moving points of light, often with luminous tails, caused by objects travelling quickly across the sky. Such shooting stars have been known since antiquity but it was not until the beginning of the nineteenth century that their true nature became fully understood.

Meteorites are not as numerous as meteors and are in entirely different kind of heavenly body. They are not merely large meteors, nor are they related to comets which they resemble visually as streaks in the sky. Rather they are considered to be much more closely associated with the asteroids or minor planets.

High-velocity particles

A meteor is a tiny particle, usually smaller than a grain of sand, moving around the Sun. It is so small that it can be seen only when it enters the Earth's upper atmosphere. With a velocity of entry possibly as high as 72 km (45 miles) per second, the meteor sets up friction with the air molecules, which causes it to destroy itself well before it reaches the ground. The resulting luminous streak in the sky, which is characteristic of the shooting star, is not produced by the meteor itself but by its effect on the atmosphere through which it is falling.

Meteors are of two main kinds: shower and sporadic. Sporadic meteors may appear from any direction at any time. Shower meteors, on the other hand, are associated with comets. The famous Leonid shower meteors of November, for example [3, 4], are linked with the faint periodical Tempel's comet and move in the same orbit as the comet itself. It has been said that meteors are mere cometary debris. This may be something of an oversimplification, but it is certainly true that one periodical comet, Biela's, was seen to disrupt and has now been replaced by a meteor shower [1, 2]. And there can be no doubt that as a comet moves along it "sheds" meteoric material.

The richness of a meteor shower is measured by its so-called zenithal hourly rate (ZHR). This is a measure of the number of meteors that would be seen by a watcher observing under ideal conditions with the shower radiant at the zenith. The most reliable annual shower, that of the Perseids, has a ZHR of about 70. Meteors below naked eye visibility are not included, so that in fact there are many more meteors than might be thought. Those that are too small to produce any luminous effects are known as micrometeors and are extremely numerous.

Because the meteors in a shower are travelling through space in parallel paths they seem to radiate from one particular point in the sky, which is known as the radiant. The principle is analogous with the view from a bridge overlooking a motorway. The parallel lanes of the motorway will seem to meet at a point near the horizon, which may be termed the apparent radiant of the lanes. Thus, on the same principle, the November Leonids have their radiant in Leo, the August Perseids in Perseus and so on.

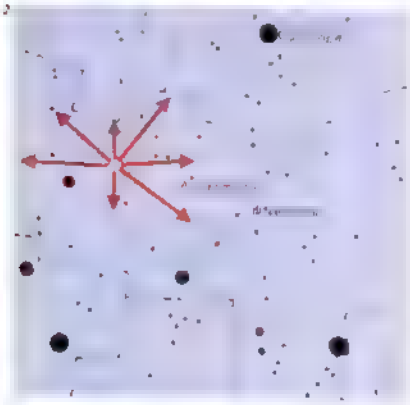
Regular annual showers

Showers of meteors occur regularly on an annual basis. These include the Quadrantids (1–6 January, maximum 3–4 January)

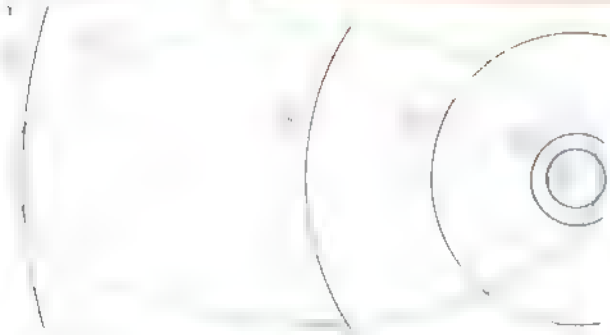
CONNECTIONS

See also

1 Biela's comet once had a period of 6.75 years. In 1846, as shown in a contemporary drawing by Angelo Secchi (1818–78), it divided into a pair of comets. The division may have been caused initially by a close approach to Jupiter in 1842 with the pull of the Sun accounting for the rest of the change. The two comets, separated by over 2 million km (1.25 million miles), returned in 1852. Because of their unfavourable positions they were not seen in 1858 and did not appear in 1866. They have not been seen again.

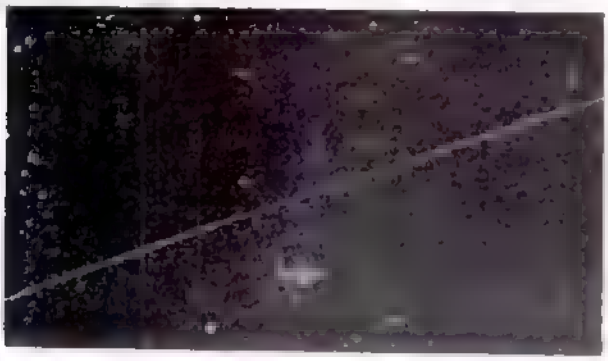


2 A brilliant meteor shower (red) from a radiant in Andromeda (blue) where Biela's comet should have been was seen in 1872. It was probably the comet's debris. The shower is now extremely feeble.



3 The orbit of the Leonid meteor stream (1) intersects the orbits of the Earth (2), Mars (3), Jupiter (4), Saturn (5) and Uranus (6). Because the meteors are not distributed evenly, major meteor showers appear only occasionally. The average interval was once just over 33 years, but the expected showers of 1899 and 1933 were missed because the shower orbit had suffered planetary perturbations.

4 A splendid Leonid meteor shower was photographed from Arizona on 17 November 1966. The shower was not visible from Europe.



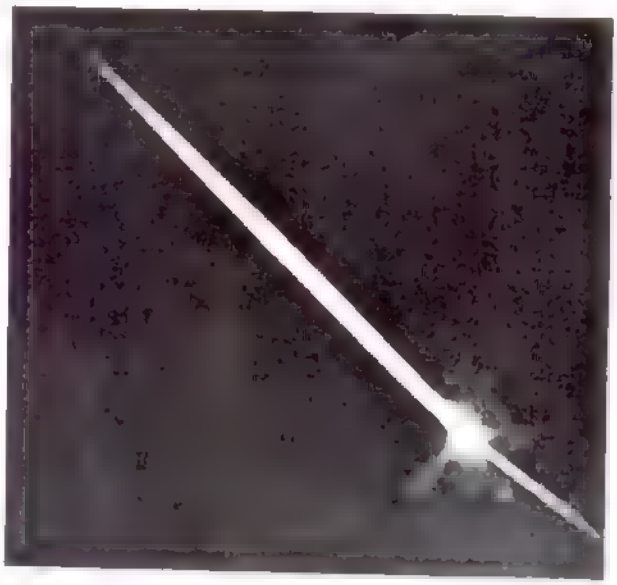
5 A meteor trail is seen here near the cluster Praesepe, the meteor was of about the second magnitude and lasted 1.5 secs.

Lyrids (19–24 April); Eta Aquarids (1–8 May – associated with the famous Halley's comet), Perseids (25 July–18 August), Orionids (16–26 October), Taurids (20 October–30 November), the far southern Phoenicids (4–5 December), the Geminids (7–15 December); and Ursids (17–24 December). The Leonids, which are at their peak on 17 November, are less reliable because the meteors are clustered rather than spread along the cometary orbit. Thus a major shower cannot be seen until the Earth passes through the main swarm. This happened in 1799, 1833, 1866 and again in 1966, so that there could possibly be another display in 1999. In the years between these major showers the Leonids are sparse.

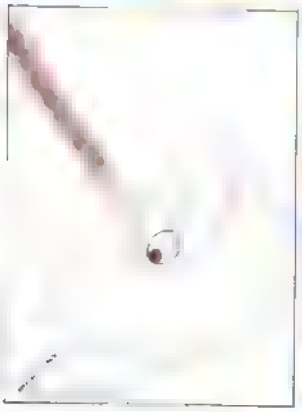
The history of meteorites

A larger body encountering the Earth may survive the journey to the ground without being destroyed. These meteorites may be of several kinds. Aerolites [10] are mainly of stone while siderites [12] have a high percentage of iron. There are various intermediate types. Etching a meteorite with acid will show the characteristic forms, known as the Widmanstätten patterns, and this is one sure way of telling whether a piece of material is of meteoritic origin or not.

Meteorites have been known for many centuries – for instance, the Sacred Stone at Mecca is certainly a meteorite – but for a long time their cosmic origin was doubted. In 1795, when a 25kg (56lb) meteorite fell at Wold Cottage in Yorkshire, it was suggested that the object was a stone hurled out of the Icelandic volcano Hekla. However, in 1803 a meteorite group fell at L'Aigle in France and the famous astronomer Jean-Baptiste Biot (1774–1862) was able to demonstrate that the objects definitely came from the sky. Most museums have meteorite collections; the largest meteorite on display, at the Hayden Planetarium in New York, weighs about 31 tonnes. Fortunately major meteorite falls are rare. The best example of craters produced by large meteorites are the Canyon Diablo Crater in Arizona [6, 7] and the Wolf Creek Crater in Australia. There is no known case, to date, of any human being having been killed by a falling meteorite.

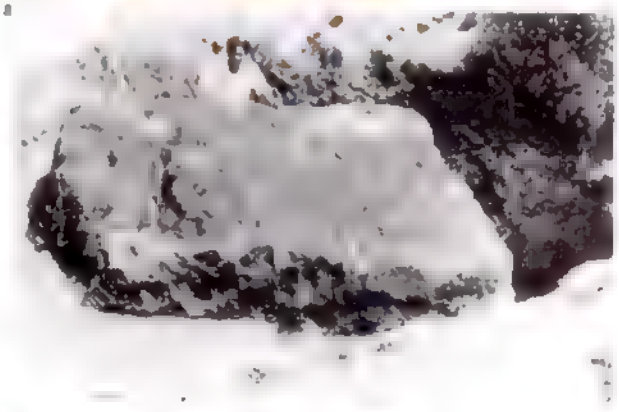


The exploding Andromedid meteor was photographed on 23 November 1895. This is one of the finest meteor photographs.

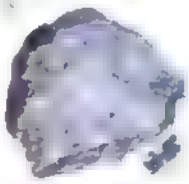


6 The Arizona meteorite crater was formed by several nickel-iron meteorites. Burning up as they plunged through the atmosphere (A), they shattered the Earth's outer layer of rock on impact (B). Because of their high speed they burrowed, causing friction, heat, compression and shock waves (C), culminating in a violent explosion (D) that left a crater. Areas of meteorite fragments (E) show those that were unaltered by heat (blue), small heat affected fragments (yellow), heavy heat altered boulders (black) and metallic spheroids formed by condensation (red).

7 The impact crater near Winslow, Arizona, is more than 1km (0.6 mile) in diameter. It may well be over 10 000 years old. Many meteorite fragments have been found in the area.

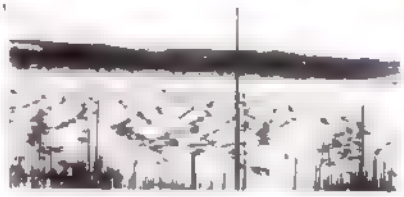


8 The Hoba West meteorite near Grootfontein in South West Africa is the largest known meteorite. It weighs over 60 tonnes and its weight before entering the Earth's atmosphere may have been 20 tonnes more. The meteorite still lies where it fell in prehistoric times. No crater was produced. A meteorite of this kind could be highly destructive but fortunately major falls are extremely rare.



9 The Orgueil meteorite (1864) is carbonaceous, having organic compounds.

10 The Norton Furnas aerolite of 1948 is the heaviest (over 1 tonne) of its type of stony meteorite.



11 The most destructive fall of modern times was that of 1908 in the Tunguska region of Siberia. A meteorite came down in forested country, flattening pine trees for several miles around.

12 A siderite (iron meteorite) when cut and etched with acid may show what are called Widmanstätten patterns, caused by a crystalline metallic structure that requires unusual conditions for formation. Such crystals are unique to meteorites.

13 Tektites, found in Australia and in some other areas of the world, are small, glass-like objects, often aerodynamically shaped. They appear to have been heated twice but their precise origin is still unknown.

The Sun and the solar spectrum

The Sun is a star, one of 100,000 million stars in our Galaxy. In the universe as a whole it is insignificant and is classed as a yellow dwarf star with a spectrum of type G, but in our planetary system – "the Solar System" – it is the all-important controlling body.

Immensely larger than the Earth, the Sun is made principally of hydrogen and helium and has a diameter of 1,392,000km (865,000 miles). Although it is big enough to contain more than a million bodies the volume of the Earth, its mass is only 1.990×10^{33} grammes – that is to say, approximately 333,000 times that of the Earth. The reason why it is not as massive as might be expected is that its density is lower than that of an Earth-type planet. The mean value for the specific gravity is 1.409 (that is to say, 1.409 times that of an equal volume of water) but the Sun is not homogeneous and the density increases rapidly beneath the brilliant outer surface.

The Sun lies some 32,000 light-years from the centre of our Galaxy and takes approximately 225 million years to complete one journey round the galactic nucleus. It has an axial rotation period of 25.4 days at the equator, but this period is considerably longer near the solar poles because the Sun does not rotate in the manner of a solid body.

The photosphere

The bright outer surface of the Sun is known as the photosphere and has a temperature of 5,500°C. On it may be seen darker patches, which are known as sunspots [Key]. These are not truly black, but appear so by contrast; if a spot could be seen shining on its own its surface brilliancy would be greater than that of an arc-lamp.

To look at the Sun through any telescope or binoculars will almost certainly blind an observer permanently and dark filters are unreliable safeguards. The only sensible method is that of projection – using a telescope to throw the solar image on to a screen held or fastened behind the eyepiece. The Sun is not as generally smooth and featureless as might be thought; granules exist on it, each of which is about 1,500km (1,000 miles) in diameter. Convection currents occur below the Sun's outer layer, and it is the rising gas columns they generate that cause the granules, whose dark edges show cooler gases dropping downwards [7].

A typical large sunspot consists of a central dark umbra surrounded by a lighter area of penumbra, although the shapes are usually very irregular and spots tend to occur in groups – generally with two main spots, one "leader" and one "follower" [3]. Some groups may be immensely complex and of tremendous area, but they are not long-lived. Even a large group will generally persist for only a few months at most, while smaller spots may last for only a few hours. As the Sun rotates the spots may be seen to be carried slowly across the disc from one side to the other. It takes about a fortnight for a spot to make the full crossing. After a similar interval, it will reappear on the opposite side of the disc – provided that it still exists.

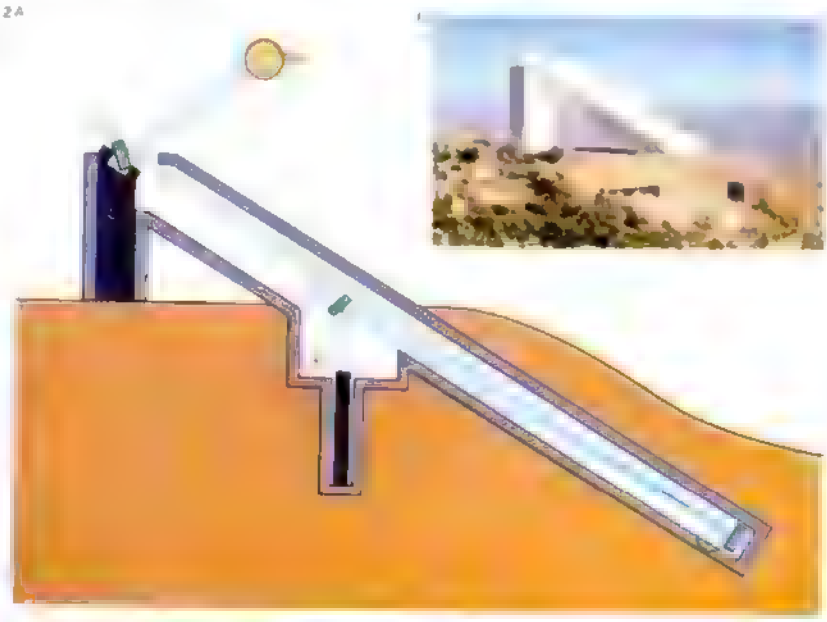
Regular cycles

The cycle of solar activity is fairly regular and has a period of 11 years, thus there were maxima in 1957-8 and again in 1969-70 when groups were plentiful [4B]. At the intervening spot-minima the disc may re

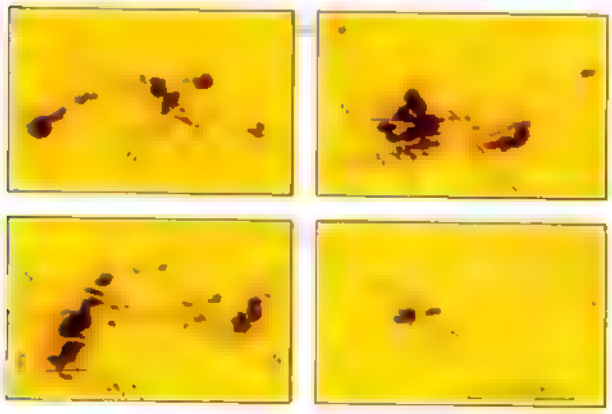
CONNECTIONS

See also

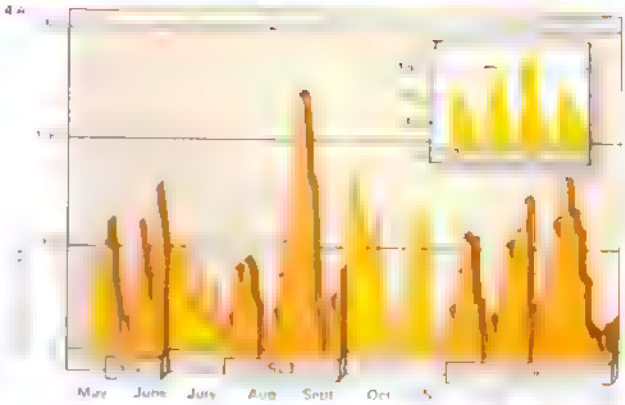
1 The coelostat, used to observe the Sun overcomes the problem of manoeuvring large telescope mounts with a movable mirror [1] that rotates with the Sun while another is stationary [2]. One axis [3] of the rotating mirror points to the celestial pole and the other is adjustable.



2 The heliostat [A], an elaborate version of the coelostat, is installed in the 500ft (152.4m) telescope at Kitt Peak, Arizona [B], to track the Sun. Sunlight falls on a rotating mirror [1]; it reflects down a tube to a concave mirror [2], focuses on to a plane mirror [3] and passes through a spectrograph [4].

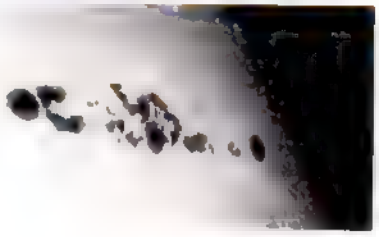


3 The sunspot observations of 1947 show that on 11 February [A] the identities of the leader and follower were still in doubt. But as the lines od- cate, the magnetic polarities were clear. From 9 March [B] to 7 April [C] the leader [1] and follower [2] are distinct. By 5 May [D] activity has ceased.



4 Sunspot numbers fluctuated widely [A] during the Skylab mission between 14 May 1973 and 8 February 1974. The periods when the Skylab was manned (28, 59 and 84 days respectively for the three crews) are shown in orange on the large graph (SL2, SL3 and SL4). The inset graph [B] charts the sunspot cycle from 1935 to 1973. The 1969-70 maximum was much less intense than that of 1957-8.

5 The penumbra of sunspots [1], near the Sun's rim (the limb) seems to narrow on the side nearest the centre of the disc, indicating that the spots are depressions in the photosphere. This effect was first noted by A. Wilson in 1769.



6 A sunspot near the limb illustrates the Wilson effect. A narrowing of the penumbra towards the centre of the Sun's disc is clear. The bright streaks are called faculae and are usually associated with major sunspot groups.

main featureless for many days on end

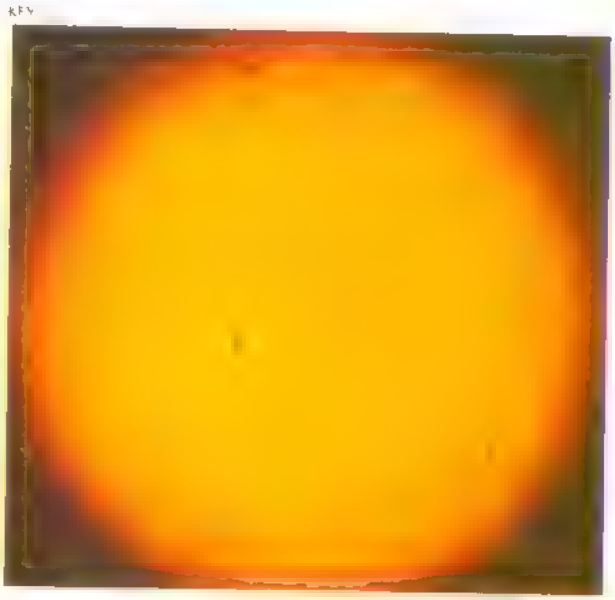
Sunspots are associated with very strong magnetic fields and this has led to the modern theory of spot formation laid down by Harold Babcock (1882 1968) in 1962 The Sun has an overall magnetic field and it may be assumed that the lines of magnetic force run from one pole to the other below the bright surface Owing to the difference in rotation period between the equatorial and the polar zones, the magnetic lines become distorted over an interval of some years and are pulled out" along the equator, while the polar magnetic field is reinforced and becomes unstable. Eventually a loop of magnetic energy breaks through the surface producing two spots, one with north polarity and the other with south Because of the magnetic linking, the polarities for the leading and following spots are opposite in the two hemispheres of the Sun After about 11 years the "knots" in the lines break and the Sun reverts abruptly to its original state But for the following cycle the polarities of spots in the two hemispheres are reversed

Visual studies of the Sun's photosphere give us only limited information to draw on and most of our knowledge comes from instruments based on the principle of the spectroscope. According to the laws of spectroscopy as laid down by G Kirchhoff (1824–87) in 1859, an incandescent solid, liquid or high-pressure gas will produce a continuous or rainbow spectrum, while a low-pressure gas will yield an emission spectrum consisting of isolated bright lines

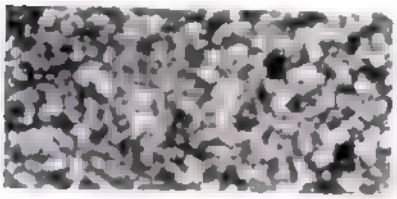
New elements found

Gaseous elements in the Sun's atmosphere absorb light at specific frequencies from the continuous spectrum emitted by the photosphere, producing gaps or dark lines on the spectrum These dark lines are called, in the case of the Sun, Fraunhofer lines. The elements present in the Sun's atmosphere can be identified from the positions (that is, the frequencies) and intensities of the lines on the spectrum

In this way more than 60 elements have already been found in the Sun One of them – helium – was even identified in the Sun before it was found on Earth

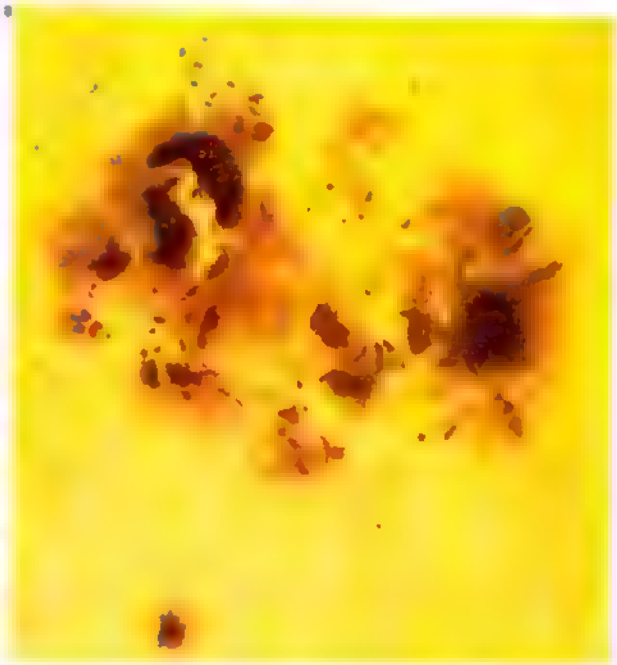


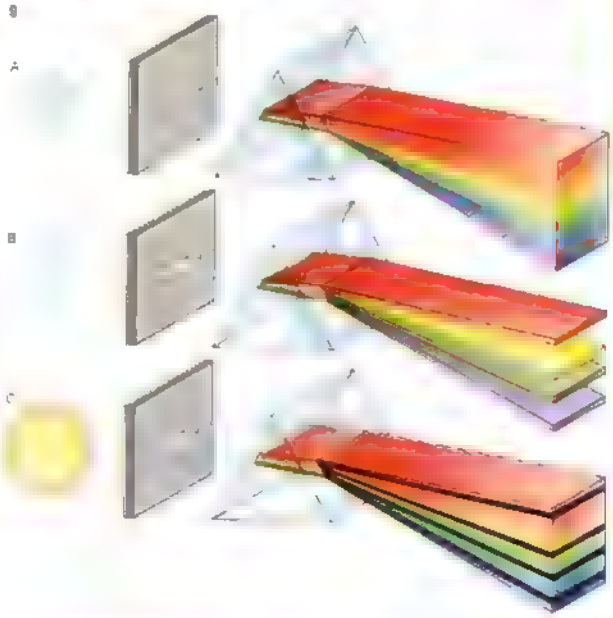
The solar maximum of 1958 was the most energetic ever recorded The photograph shows a heavily spotted disc



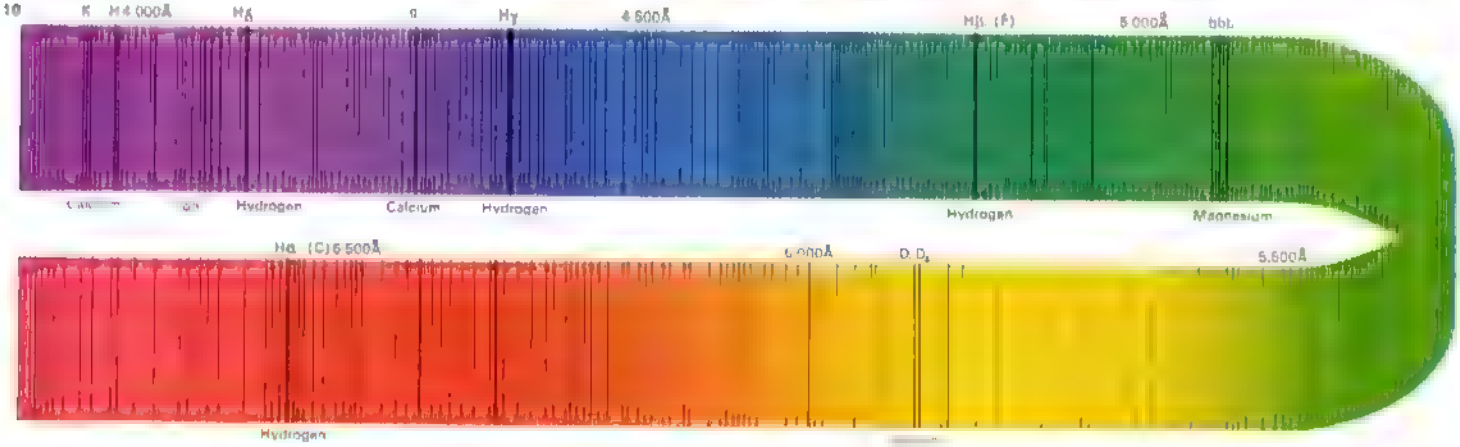
7 Solar granulation results from convection effects These gaseous columns occur all over the Sun's disc Their average diameter is about 1,500km although their size range is fairly wide

8 Large and complex sunspot groups were still common when the Sun was well past the peak of its 1947 cycle of activity This group was photographed at Mount Wilson on 17 May 1951





9 The solar spectrum combines two effects The photosphere, like a gas under high pressure in the laboratory produces a rainbow or continuous spectrum from red at the long-wave end to violet at the short wave end (A) The solar atmosphere, like gas under lower pressure should in theory produce an emission spectrum (B) consisting of isolated bright lines, each associated with a particular element However, as light is radiated from the Sun's surface, gaseous elements in the solar atmosphere absorb specific wave lengths so that the spectrum observed on Earth has gaps (dark lines) – the Fraunhofer lines – in it



10 Lines in the solar spectrum can be accurately mapped with the 13ft (4m) spectrograph at the Mount Wilson observatory The range illustrated is from 3 900 to 6 900 ångstroms, ie from violet through to red (One ångstrom equals one hundred millionth of a centimetre) Each line can be identified thus the D lines in the centre of the spectrum are due to sodium the H Alpha line to hydrogen

The Sun's atmosphere and radiation

Ordinary telescopes show only the bright surface or photosphere of the Sun and features such as the spots, the granulation and the bright faculae (temporary patches on the Sun's surface) which lie above the photosphere itself. More complicated methods are needed to study the solar atmosphere because the Sun's surroundings can be seen with the naked eye (or with an ordinary telescope) only during the rare intervals when the Moon covers the Sun completely and produces a total eclipse.

Prominences and flares

The part of the solar atmosphere lying immediately above the photosphere is called the chromosphere ("colour sphere") because it has a characteristically reddish appearance. This is also the region of the [illegible] brilliant prominences. To observe the prominences, instruments based on the principle of the spectroscope are used. There are two main types of prominences, eruptive [2] and quiescent [Key]. Eruptive prominences are in violent motion and have been observed extending to more than 500,000km (310,500 miles) above the Sun's surface. Quiescent prominences are much more stable and may hang in the chromosphere for days before breaking up. Both are most common near the peak of the solar cycle of activity.

Prominences are often associated with major spot-groups. Active groups also produce "flares", which are not usually visible, although a few have been seen. The flares are short lived and emit streams of particles as well as short-wave radiation. These emissions have marked effects upon the Earth, producing magnetic storms or disturbances of the Earth's magnetic field that affect radio communications and compasses. They also produce the beautiful solar lights or aurorae [3, 4]. The Sun also sends out a constant stream of low-energy particles in all directions, making up what is now known as the solar wind. It is this emission that has a strong effect upon the tails of comets, forcing them to point away from the Sun.

In addition to sending out light, the Sun is an important source of infra-red (heat) and ultra-violet radiation, as well as radio waves, X-rays and gamma-rays. Studies are difficult to carry out from Earth because of the screening effect of the atmosphere, but knowledge has been greatly increased as a result of work carried out by satellites and by the Skylab astronauts in 1973–4. It was fortunate that the Sun was reasonably active [5, 6] while the astronauts were in orbit, for many of the results could not possibly have been duplicated in ground-based observations.

The powerhouse

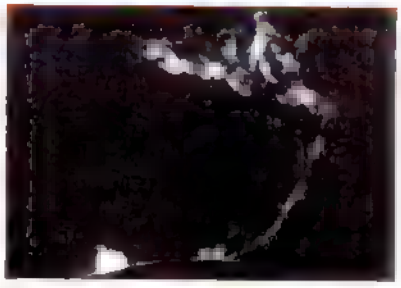
Although astronomers cannot prove most of their theories about the nature of the Sun they have a good idea of its composition. The temperature increases toward the core, until at the centre of the globe it is estimated at about 10 million degrees centigrade. It is here, in what is called the Sun's "power house", that the energy is being generated.

It is erroneous to suppose that the Sun is burning in the same way that a fire burns. A Sun made up entirely of coal, and radiating as fiercely as the real Sun does, would not last long on the cosmic scale, and astronomers believe that the Sun is at least 5,000 million years old. (It is certainly older than the Earth

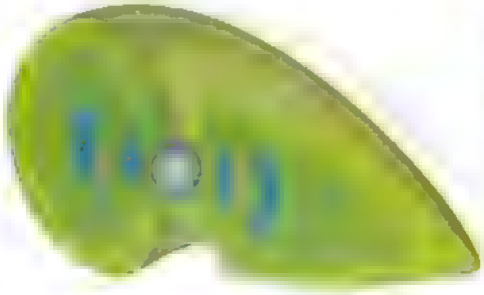
CONNECTIONS

See also

1 The main structure of the Sun cannot be drawn to an accurate scale. In the solar interior [1] nuclear transformations create energy. The convective zone [2] leads out to the relatively rarefied photosphere [3] which is surprisingly narrow and has sharp boundaries. Spots [4] lie in the photosphere and associated with them are the Flares [5] and the prominences, which lie in the chromosphere [6]. The temperature of the chromosphere rises from 6,000°C at the bottom to more than 50,000°C near the upper portion. (Temperature here is purely a measure of the speeds at which the atomic particles are moving and does not indicate extra "heat".) In the chromosphere there are spicules [7] masses of high-temperature gases shooting up rapidly into the immensely rarefied corona [8]. The corona is large and streamers [9] issue from it.



2 A large eruptive prominence occurred on 4 June 1946 at 16.03 hours [A]. It took the form of an arch. By 17.03 hours [B], it had been blown to 322,000 km (200,000 miles) above the Sun's surface. Little remained of the great arch by 17.23 hours [C] the prominence is shown dispersing.



3 Aurorae frequently associated with flares are due to charged particles sent out by the Sun. The particles come from the Sun [1] and enter the Van Allen belts [2] which surround the Earth [3]. The Van Allen belts become overloaded and particles cascade down into the upper air producing the aurora glows.

4 Aurorae may take many forms such as curtains, arches and draperies in with varied and lovely colours. They are aptly called "flaming surges". The electrified particles tend to spiral towards the Earth's magnetic poles, which is why aurorae are best seen from high latitudes. In low latitudes they are rare. Aurorae are commonest when the Sun is active, near the maximum of its 11-year cycle.

which has an age of about 4,600 million years.) The source of solar energy is to be found in nuclear transformations. Hydrogen is the main constituent and near the core where the temperatures and pressures are so extreme, the second lightest element, helium is formed from hydrogen nuclei by nuclear fusion. It takes four hydrogen nuclei to make one nucleus of helium, in the process a little mass is lost, being converted into a large amount of energy. The energy produced keeps the Sun radiating: the loss of mass amounts to four million tonnes per second. This may seem a lot but it is negligible compared with the total mass of the Sun; there is enough hydrogen available to keep the Sun shining in its present form for at least another 5,000 million years, perhaps longer.

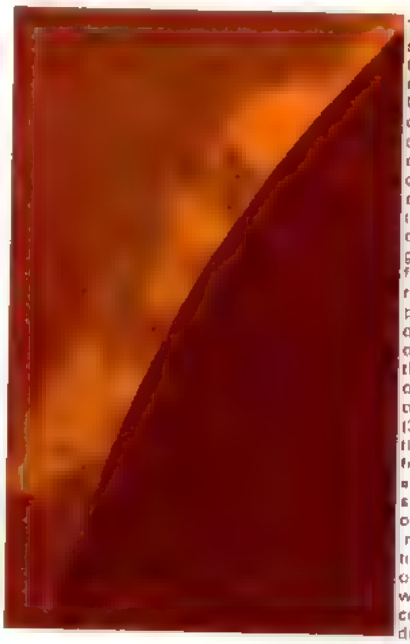
Eventually the hydrogen will start to become exhausted and the Sun will change its structure drastically. According to present theory, it will pass through a red giant stage when it will have a luminosity at least 100 times as great as it is today; it will then collapse into a small dense star of the type known as a white dwarf.

The Earth also has a limited life-span. It cannot survive the red giant stage; along with the other inner planets it will certainly be destroyed eventually.

Solar research

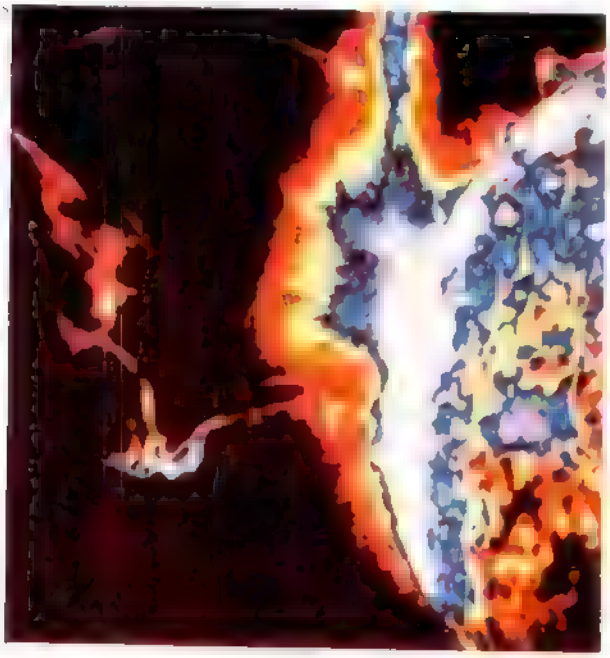
Knowledge of the Sun has been drawn from many different areas of research. Radio astronomy is of especial importance. This is a method of studying astronomy in the long-wavelength region of the electromagnetic spectrum. The Sun is a strong radio source, a fact known since the early days of radio astronomy. The study of X-rays and gamma rays from the Sun is much more recent because it depends upon instruments operating from above the Earth's hampering atmospheric layers.

There has been much discussion about the extent of the solar atmosphere. Beyond the chromosphere lies the corona, which is immensely rarefied and has no definite boundary. It merely "thins out" to become the solar wind. The Sun is a much more complicated and varied body than was originally thought, but it is probably a typical sta.



The most spectacular solar features are the streams of hot gas called prominences. Quiescent prominences may hang in the chromosphere for days or weeks, bulging out from the surface about 32 000km (20 000 miles). This example was photographed with a Lyot filter on a 4in (10cm) refractor. Eruptive prominences, the other main type of prominence, are thin flames of gas often reaching heights of 400 000km (312 500 miles); they are formed more frequently in those areas containing sun-spots. The invention of the coronagraph in 1930 enabled continuous photography of prominences which otherwise could be seen only during a total eclipse.

5 This solar prominence was photographed by the astronauts on board Skylab. The colours, in this extreme ultra-violet shot, are false. They represent the degree of radiation intensity from red, through yellow and blue, to purple and white where the activity is most intense. This picture could be taken only with equipment carried above the layers of the Earth's atmosphere.



6 This Skylab photograph depicts an eruptive solar prominence which is seen rising to a great height. Matter at the apex of the arch seems to be reflected back to the Sun's surface.

Solar eclipses

The Skylab missions in 1973 and 1974 considerably improved man's knowledge of the Sun because they made possible extended observations of features not visible from the Earth's surface. Before the missions the best views of those features had been obtained during total eclipses.

Features of eclipses

The Moon is so much closer to the Earth than the Sun that despite its small size it looks just as big in our sky. The coincidence is fortunate: it means that when the three bodies are exactly lined up, the Moon can just blot out the Sun's brilliant photosphere, leaving the chromosphere and the corona to shine out unhindered [1B]. The spectacle is always brief, however, because the Moon's shadow only just touches the Earth [1A]: the track of totality can never be more than 269km (167 miles) wide and the maximum duration in any one spot is less than eight minutes. Hence astronomers have always done their best to take advantage of these opportunities. It was only after the famous eclipse of 1842 that most astronomers came to believe that the prominences belonged to the Sun, rather than to the Moon.

Because the Moon has an orbit that is not circular, its apparent size varies. At apogee (the point of greatest distance from the Earth), the full Moon looks ten per cent smaller than at perigee (the point nearest the Earth). The apparent diameter of the Sun also changes, being greatest in December and least in June because of the Earth's varying distance from the Sun. If the Moon appears smaller than the Sun, it is unable to cover the photosphere completely and the result is an annular (ring-shaped) eclipse, leaving a ring of sunlight showing round the dark mass of the Moon [1D]. There are also partial eclipses [1C], when not all of the Sun is hidden. Annular and partial eclipses are relatively unimportant because the Sun's surroundings do not come into view.

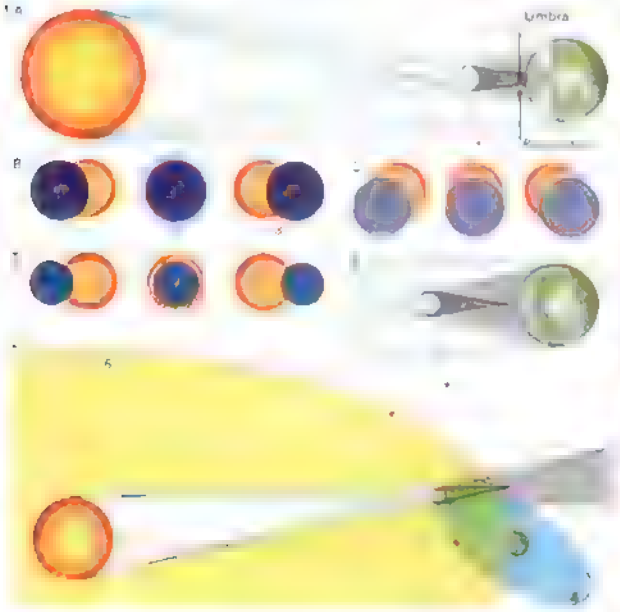
Eclipse records date back many centuries; there are records of an eclipse seen from China as long ago as 2136 BC. Eclipses do not occur at every new Moon because the lunar orbit is appreciably inclined to that of the Earth [1F]. Any eclipse, however, may be followed by another similar eclipse (total or partial) 18 years 10.3 days or 11.3 days with five leap years later when the Sun, Moon and Earth then return to almost the same relative positions. This period is known as the Saros (although there will be other eclipses in the intervening period). The Saros [Key] is not exact, but it is better than nothing at all as far as predictions are concerned and the ancients made extensive use of it.

Observing the corona

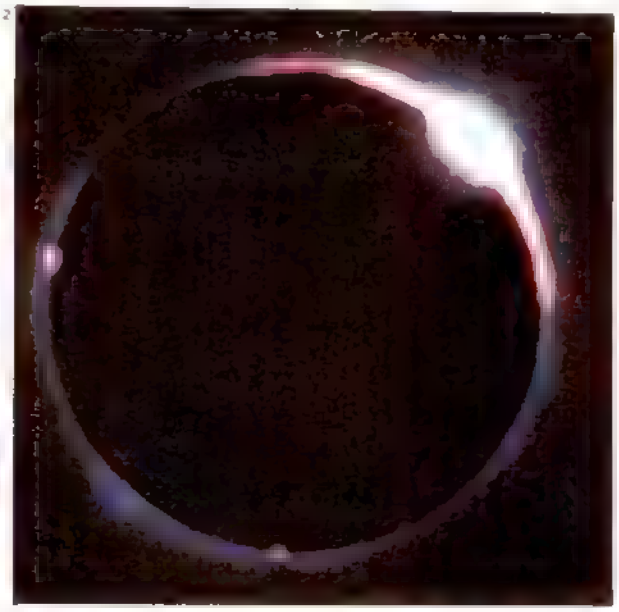
The main glory of a total eclipse lies in the view of the corona [5, 6]. It has been found that the shape of the corona varies according to the state of the solar cycle. Near spot-minimum the corona is fairly symmetrical, whereas near spot-maximum there are long streamers. The sky is dark enough for planets and bright stars to be seen and on several occasions unexpected comets have been found close to the hidden Sun. It is a pity that total eclipses are so rare in any particular place on Earth. In England, for example, the last total eclipse occurred in 1927; the next will not be until 1999.

CONNECTIONS

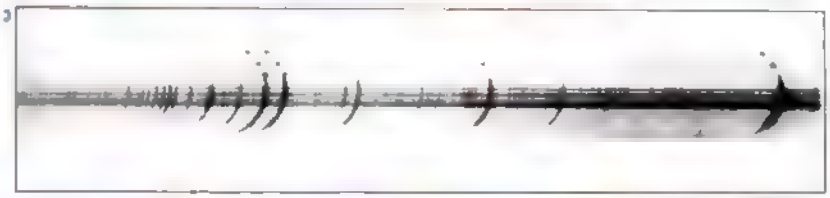
See also



1 In a total eclipse (A) the main cone of shadow or umbra (not to be confused with the umbra of a sunspot) reaches the Earth's surface while to either side of it, in the partial shadow or penumbra, a partial eclipse is seen. The Moon and Sun (B) approach totality (1), arrive at totality (2) and leave totality (3). C shows a partial eclipse that is not total anywhere on Earth. An annular eclipse (D) occurs when the umbra stops short of the Earth. E, F shows how the tilt of the Moon's orbit (4) to the plane of the Earth's orbit (5) prevents an eclipse from occurring every month of the year.



2 As the Sun's disc reappears from behind the Moon after totality there is a glorious "diamond ring" effect, as occurred in the total eclipse of 21 November 1966. It lasts for only a few seconds.



3 Just before totality begins, or just after it ends, at the moment of the "diamond ring" effect, the Sun's atmosphere is seen without the background photosphere. The dark absorption lines become suddenly bright emission lines, producing the "flash spectrum", a negative of which is shown here. The effect is brief, but it has been photographed many times and much information, particularly about the solar atmosphere, has been acquired.

4 During the partial eclipse of 21 Nov 1966 several sunspots were seen on the disc.

5 Photographing an eclipse from an aircraft has advantages – there are no clouds and the aircraft can follow the Moon's shadow. Here the chromosphere and inner corona are clear.

6 The 1961 eclipse, taken with a long exposure, shows more of the outer corona extending over a wide area. The inner corona and prominences are overexposed as a result.

With the development of spectroscopy as a research method, it became possible to study the chromosphere and the prominences at any time. The corona, however, is a more difficult feature to investigate because even its inner part is much fainter than the chromosphere. The French astronomer Bernard Lyot (1897–1952) developed a device known as a coronagraph, which can be used from high-altitude observatories to observe the inner corona. From Earth, the outer part remains unobservable except during totality. Knowledge of the Sun is therefore incomplete and attempts to increase this knowledge are not helped by the fact that certain radiations in the electromagnetic spectrum, including X-rays, can never penetrate to the surface of the Earth.

In the past, various methods have been worked out to overcome the screening effects of the Earth's atmosphere. For instance, balloons have been used, although they cannot fly as high as astronomers would like. The final solution was achieved when sophisticated equipment for studying solar X-rays was taken up in Skylab and used to good effect [7]. The corona was examined at all wavelengths and it is probably true to say that future total eclipses will be less important than those of the past. The first Skylab crew returned with exposed film of the corona, representing more hours of observations than had been acquired in the millennia of man's observations during natural eclipses.

Future studies of the Sun

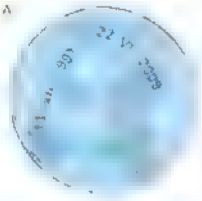
Despite the newly acquired knowledge many problems remain. A more complex coronal structure than was previously known has been revealed through Skylab ultraviolet experiments. X-ray pictures have also revealed low-density coronal regions, or coronal holes, which could be the source of disturbance in the solar wind.

Undoubtedly the future of the study of the Sun's outer surroundings lies in space-research methods. Skylab has shown the way and future orbiting stations, as well as equipment operated from the surface of the Moon will tell man much more about his own particular star than he can claim to know at the present time.

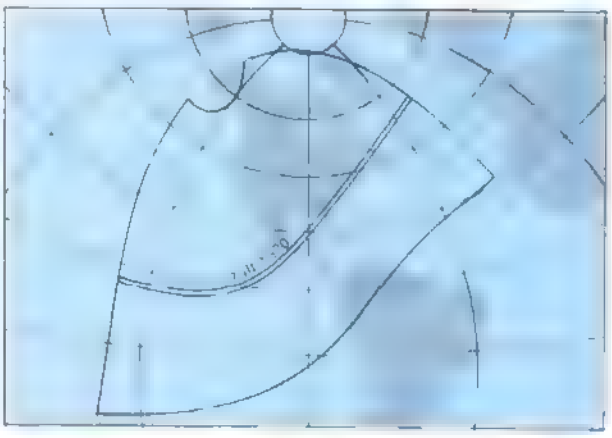
KEY

The 12 solar eclipses whose tracks of totality are shown here (A) belong to the same Saros family. The eclipse of 1991 will be total over Central America, but the next one in the cycle will be only partial: there that is, the returns are not exact. The eclipse of 7 March 1970 (B) was total over Mexico and Florida but the partial eclipse covered almost all North America. This made observation convenient.

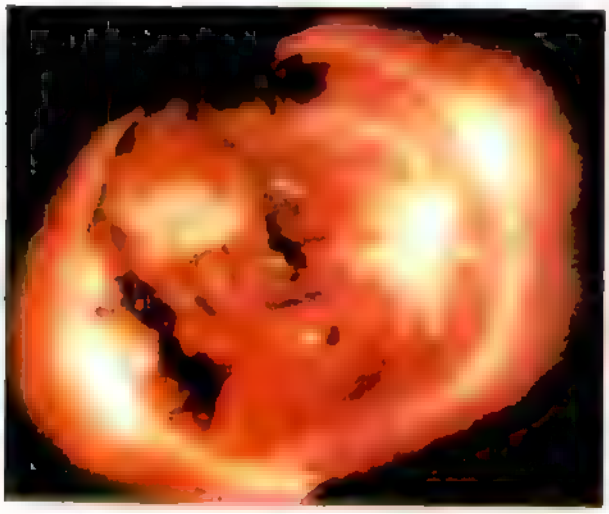
A



B

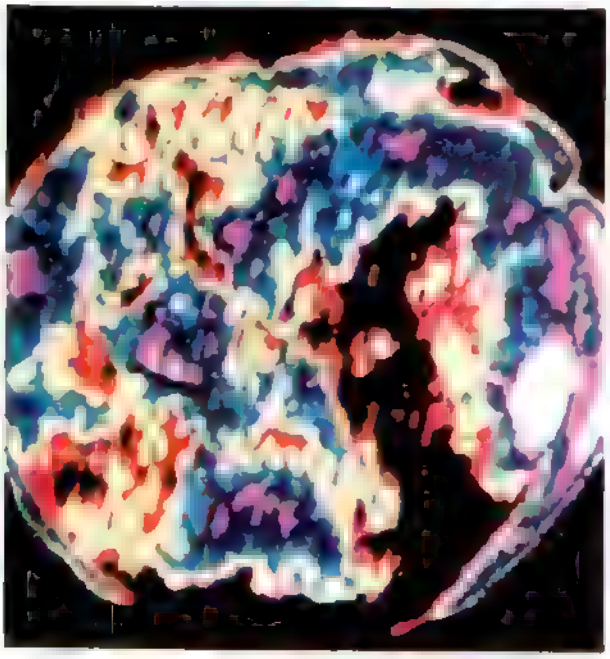


7

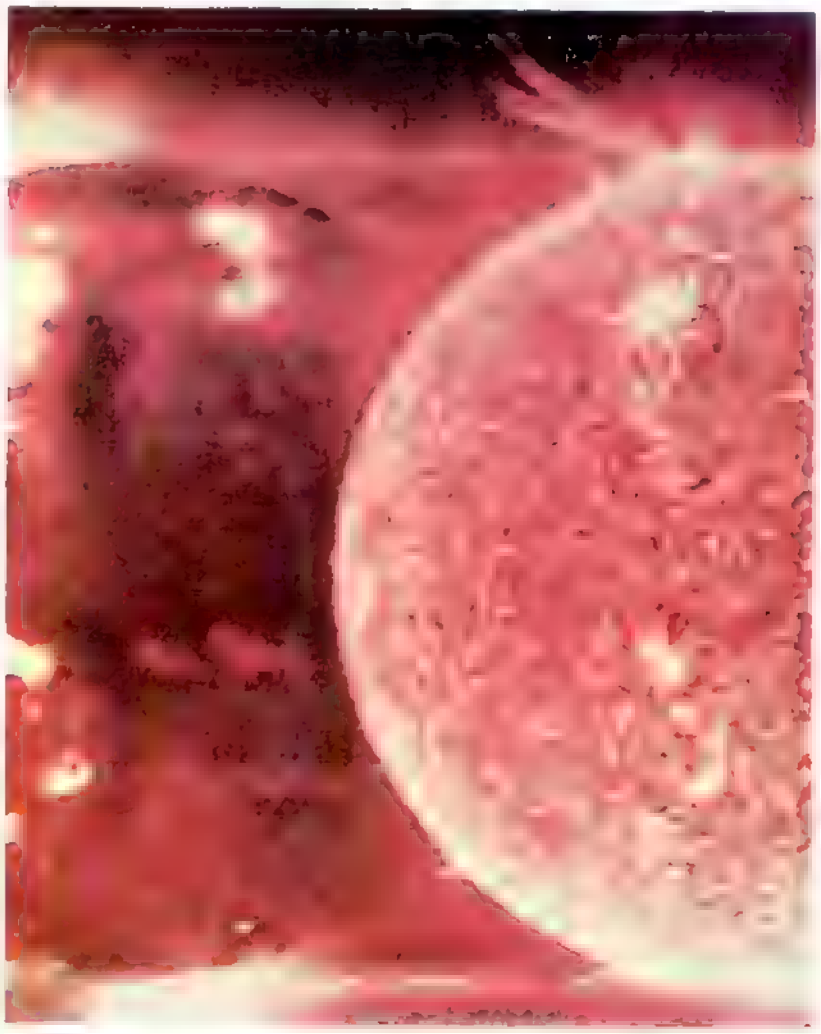


7 This X-ray image of the Sun's corona was taken by Skylab at the total eclipse of 30 June 1973. The Sun was near the minimum of its cycle of activity, and the corona was fairly symmetrical. The dark lane near the top is a "corona hole".

8



9



8 A large coronal hole is indicated by the dark stream in this "false colour" extreme ultra violet picture of the Sun taken from Skylab. It illustrates that the structure of the corona itself is far from being perfectly uniform.

9 One of the most impressive events recorded during the studies of the Sun carried out from Skylab is shown in this picture. Film taken on 10 June 1973 showed a huge "blob" of tenuous material the size of the Sun. It was moving outwards through the corona at a velocity approaching 400km (250 miles) per second. The size can be appreciated from a comparison with the solar disc itself. During the mission more than 40 similar event were studied but not all of them were the same size. Although the sizes of the "blobs" were so great, the actual amount of material involved was relatively slight because the material was so rarefied. It was the first time this type of event had been observed.

Star types

No star, apart from the Sun, is close enough to Earth to appear as anything but a point of light, studies of the Sun are therefore all important in stellar astronomy Moreover, the value of the telescope on its own is limited. Astrophysics mainly relies on instruments based on the principle of the spectroscope, which splits up light and gives information about the substances present in the light source

Stellar spectra

The spectrum of the Sun was first studied by Isaac Newton (1642–1727) in 1666, but real progress was not made until the nineteenth century, mainly by Joseph von Fraunhofer (1787–1826), who mapped the dark absorption lines in the solar spectrum, still often called Fraunhofer lines. The lines were correctly interpreted by Gustav Kirchhoff (1824–87) and Robert Bunsen (1811–99) in 1859 Stellar spectroscopy, however, was a much more difficult matter, because so little light was available and spectroscopic equipment had to be used in conjunction with powerful telescopes

Pioneer work, carried out largely by Angelo Secchi (1818–78) in Italy and William Huggins (1824–1910) in England, established that the stars may be divided into several reasonably well-defined spectral types. The system now adopted [3] is that drawn up at the Harvard College Observatory under the direction of Edward Pickering (1846–1919) The spectral types are given letters of the alphabet. In order of decreasing surface temperature the six main types are B, A, F, G, K and M, the complete sequence includes five more groups of rarer type, W, O, R, N and S, which denote somewhat different spectral characteristics. The sequence is alphabetically chaotic because there were several major revisions during the research period, types C and D, for instance were found to be unnecessary

The colour of a star is a key to its spectral type. Stars of types O, B and A are white or bluish-white, F and G, yellow, K orange, and the rest orange-red Subdivisions are given by figures, thus G0 is the hottest G-type star, G5 is midway in the sequence between G and K, and G9 is only slightly hotter than K0 (the exact spectral class of the Sun is G2). Conventionally stars at the beginning of the sequence are referred to as "early" type stars and those near the end (types K, M, R N and S) "late" type, although the Harvard sequence is no longer thought to be truly evolutionary. The situation is far more complicated than was once thought

The Hertzsprung–Russell classification

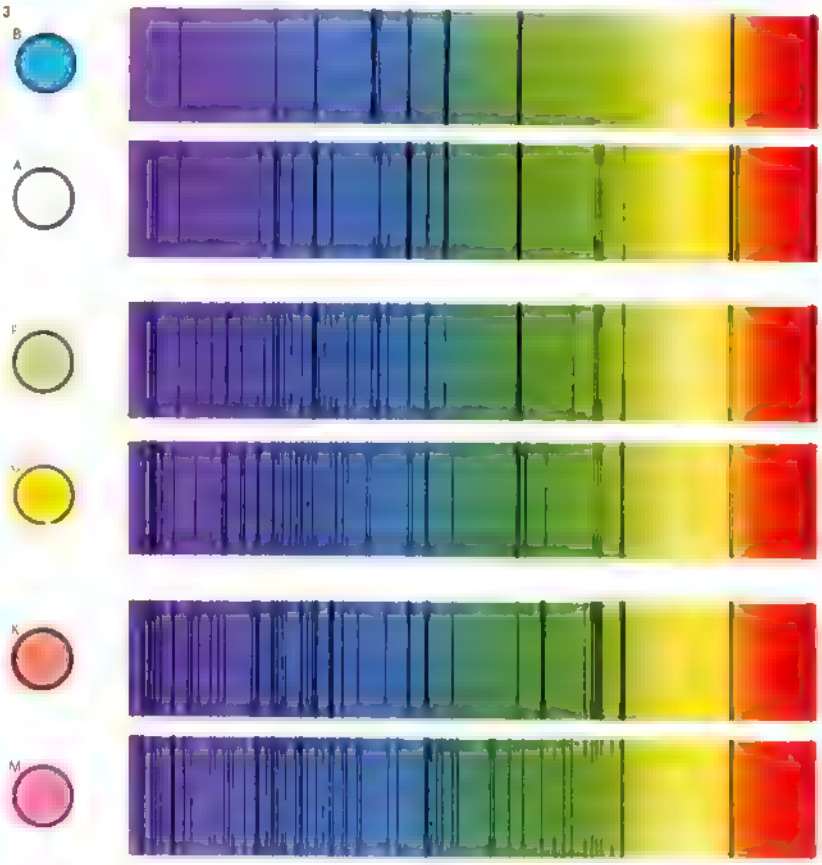
In 1908 the Danish astronomer Ejnar Hertzsprung (1873–1967) drew up a diagram in which he plotted the stars according to their luminosities and their spectral types. Research of similar kind was being carried out in the United States by Henry Russell (1877–1957) and the diagrams produced are now known as Hertzsprung Russell or H–R, diagrams [6] They have proved to be immensely informative. Even a casual glance at an H–R diagram is enough to show that the stars are not randomly distributed all over it, although the H–R diagram does not, as was once supposed, mark a strict evolutionary sequence Most of the stars in the diagram lie in a well

CONNECTIONS

See also

1 If a camera is pointed at the night sky and a time exposure made without the camera being moved, stars will appear as trails because of the rotation of the Earth The longer the exposure the longer the trail By making long exposures such as the one shown, the different colours of the stars are more easily resolved The hotter stars will appear blue or white, cooler stars yellow and still cooler stars as red trails

2 The famous constellation Crux (the Southern Cross) was photographed from Rhodesia by J McBain. The camera was attached to a driven telescope, so that the stars are shown as hard points and not as trails Of the four main stars of the Cross, three are hot and white but the fourth – Gamma Crucis – is a red giant and its colour is clearly shown here. The colours can be seen well through binoculars



3 The Harvard classification of spectral type is illustrated for the six principal classes of stars B, A, F, G, K and M The spectrum for each category is shown together with the colour symbol that is repeated on the Hertzsprung–Russell diagram opposite An example of a star in each class is given B-type stars (Rigel) helium lines are prominent 25 000°C surface temperature A-type stars (Sirius) hydrogen lines are prominent 10 000°C surface temperature F-type stars (Polaris) calcium lines are prominent, 7 500°C surface temperature The giant and dwarf division is appearing G-type stars (the Sun), 6,700°C (giants) and 6 400°C (dwarfs) surface temperature The giant and dwarf division is clear K-type stars (Arcturus) 4,100°C (giants) and 5 100°C (dwarfs) surface temperature M-type stars (Betelgeuse 3 100 C giants) and 3 500 C (dwarfs) surface temperatures many are variable and advanced in evolution

4 Magnesium lines in the green region of the Sun's spectrum were photographed by H R Hatfield with his spectrohelioscope These magnesium lines lie at a wavelength of 5,170 Ångströms

5 The double D line of sodium, also in the solar spectrum, was photographed by H R Hatfield These are among the most prominent of all spectral lines and are well shown in many stellar spectra

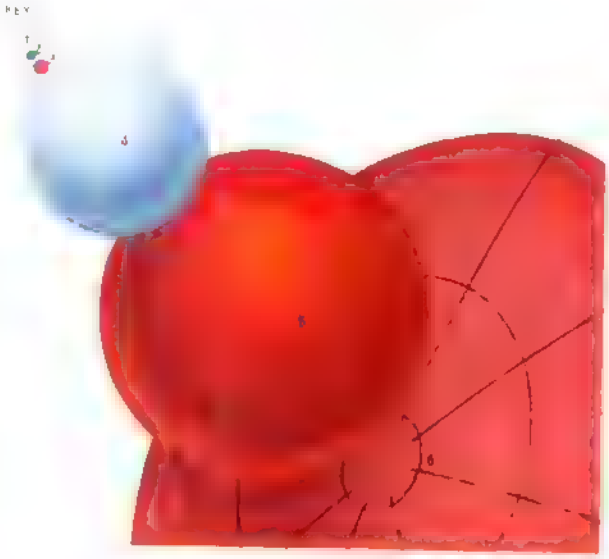
defined belt extending from the upper left to the lower right, this belt has become known as the main sequence, the Sun is a typical main sequence star.

It is also obvious that with the red and orange stars, and to a lesser extent with the yellow – that is to say, from types G to the end of the sequence – there is a sharp division into giants and dwarfs. Consider, for instance, two stars of type M: Betelgeuse in Orion and our nearest stellar neighbour Proxima Centauri. The surface temperatures are much the same, but this is the only point of similarity. Betelgeuse has a variable diameter of 420–560 million kilometres (260–350 million miles) – large enough to hold the entire orbit of the Earth – and a luminosity more than 10,000 times that of the Sun. The diameter of Proxima is less than one million kilometres (about 620,000 miles) and it has only one ten-thousandth of the Sun's luminosity (Key). M-type stars with luminosities about the same as that of the Sun do not exist, as the H–R diagram shows. The discrepancy between giant and dwarf is rather less with earlier spectral types and beyond type F it is more difficult to distinguish. (The white dwarfs, shown at the lower left of the H–R diagram, come into an entirely different category.)

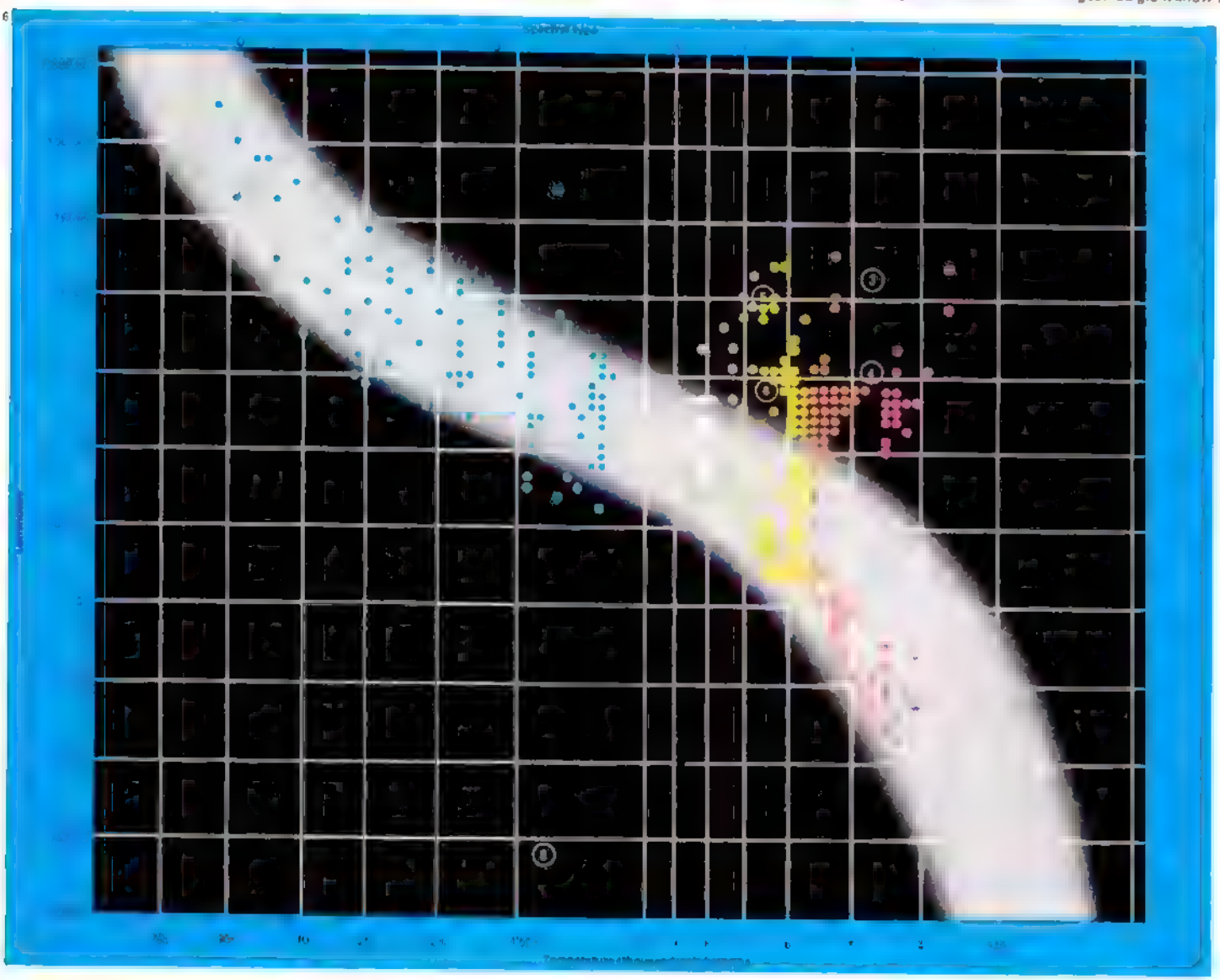
The rarer classifications

Most of the stars lie in that part of the Harvard sequence from B to M. Stars of type W have high surface temperatures – of the order of 80,000°C – and their spectra show bright emission lines, produced in the star's gaseous atmosphere. W-stars, also known as Wolf-Rayet stars, are rare, about 150 are known in our Galaxy and another 50 in the Large Magellanic Cloud. Allied to them are the O type stars, with lower surface temperatures (about 35,000°C) and both bright and dark spectral lines. Zeta Orionis or Alnitak, for example, in Orion's belt, is of type O9.

At the other end of the sequence come stars of types R, N and S. All of them are remote, so that they appear faint, and almost all are variable. They are often called carbon stars because lines caused by molecules containing carbon are so prominent in their spectra. The reddest are of type S.



Stars vary in size, temperature and luminosity. Wolf 359 (1) is a faint red dwarf and Epsilon Eridani (2) is smaller and cooler than the Sun (3) and Rigel (4) is 50,000 times brighter. Aldebaran (5) is a red giant and Antares (6) is the largest red giant known.



6 The Hertzsprung–Russell or H–R diagram is of fundamental importance. The stars are plotted on the graph according to their luminosities relative to that of the Sun, their spectral types and their surface temperatures. Most of the stars lie along the well defined belt known as the main sequence. The main sequence extends from the upper left, with extremely hot O-type stars (1), through to G-type stars (2) such as the Sun, and the red dwarfs of type M (7) of low luminosity. To the upper right lie the red supergiants (3) and the giant branch (4). Also shown are the Cepheid variables (5) and the RR Lyrae variables (6). To the lower left of the diagram are the white dwarfs (8). Stars of types K and M are divided into giant and dwarf groups which can be clearly seen.

Stellar evolution

In the early years of the twentieth century, many astronomers assumed that stars evolved strictly along the course plotted on a Hertzsprung–Russell diagram [Key], starting as luminous white stars and ending as dim red ones. According to this theory, a star would begin by condensing out of interstellar dust and gas. Gravitational forces would shrink it so that the interior would heat up. The star would start shining as a large, diffuse red giant of type M1. It would continue to contract and heat up until it joined the top of the main sequence, moving down the main sequence until it became a faint M-type red dwarf. Eventually it would be transformed into a cold, dead globe.

Evolution of a star of solar mass

That plausible sequence of stellar evolution is now known to be completely wrong. A red giant such as Betelgeuse is not young. It is very old, has used up most of its energy reserves and is in an advanced stage of evolution. The stars are now known to shine because of nuclear reactions taking place inside them and the course of stellar evolution is known to depend largely upon the initial mass of a star when it is formed from the nebular material: a massive star evolves differently from a star of much lesser mass. The only common factor is that all stars begin their careers in gaseous nebulae of which the Orion Nebula M42 is unquestionably the best-known example.

As an embryo star shrinks it heats up, but if the mass is extremely low no nuclear reactions are able to start and the star never joins the main sequence. Instead it radiates feebly until its energy has been dissipated. In the case of a star with a mass about that of the Sun [3], a stage is reached, as the gravitational shrinkage continues, when heat is carried from the interior to the surface by convection. In a short time (perhaps only a century or so) the star becomes from 100 to 1,000 times as luminous as the Sun is today. After this initial burst of glory it continues to shrink and also becomes fainter – it is approaching the main sequence. Then when the core temperature has risen sufficiently, nuclear reactions begin. Hydrogen nuclei combine to form helium nuclei, resulting in a loss of mass and the release of energy, and the star settles down on the main sequence to a long period of stable existence, lasting for perhaps as long as 10,000 million years. (The Sun, which is about 5,000 million years old, has thus reached the half-way stage in its main sequence career.)

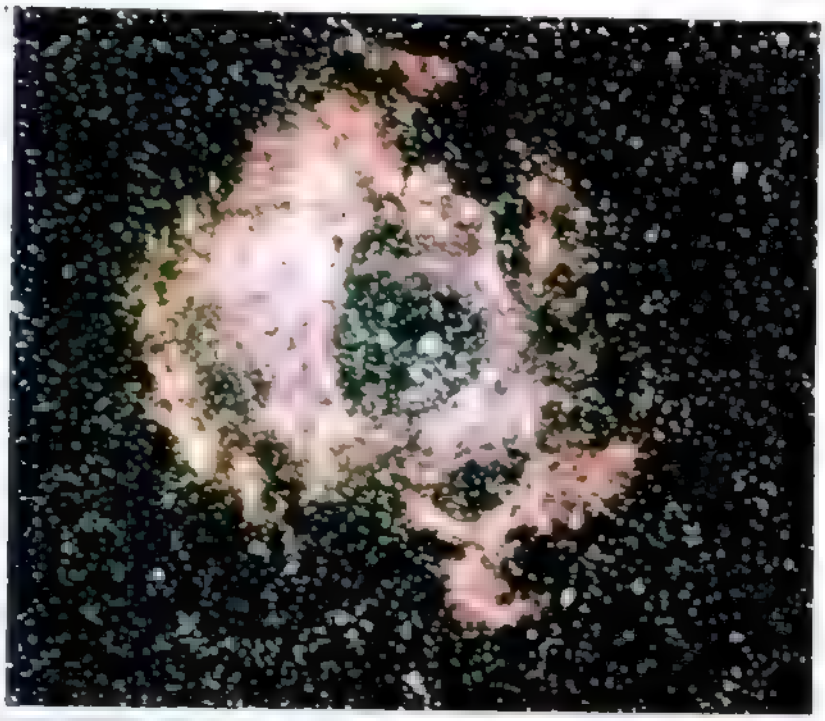
Eventually the supply of available hydrogen "fuel" begins to run low and the star has to rearrange itself. The helium core contracts rapidly and is heated up once more, enabling hydrogen nuclei to "burn" in a shell surrounding the core while the outer layers expand and cool. The star swells out to become a red giant. The central temperature rises to about 100 million degrees centigrade, although the outer layers are cool and extremely rarefied.

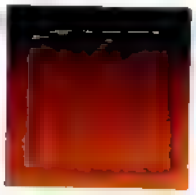
White and black dwarfs

Further types of reactions follow, but at last there is no nuclear energy left and the star collapses into a small, dense white dwarf. Because the component atoms are crushed and broken they can be tightly packed and the star's density may reach more than

CONNECTIONS

See also

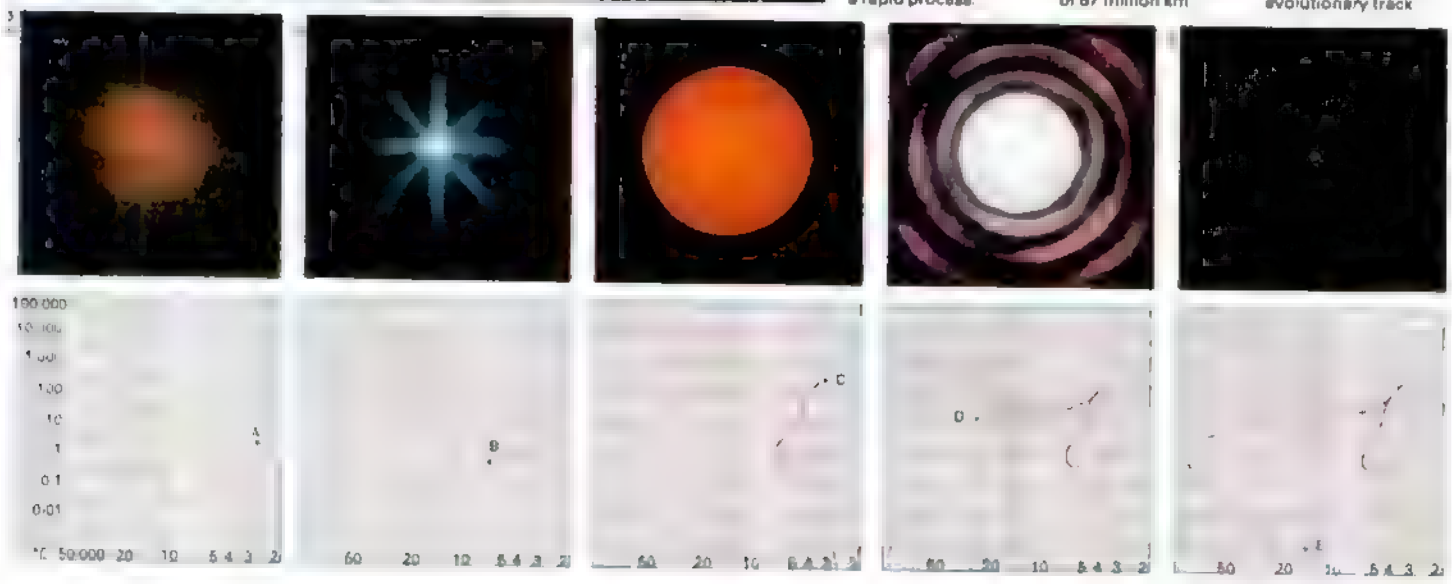




1 The Rosette Nebula, NGC 2237, is shown here in a photograph taken with the 48in (122cm) Schmidt telescope at Palomar, USA. The nebula lies in the constellation of Monoceros but is not a bright object. It is a typical emission nebula, the brightest star in it has a spectral type of O9. There is no doubt the nebula represents a region where fresh stars are being formed. Although the gas is so prominent, it is extremely rarefied and the formation of new stars within the region is by no means a rapid process.

2 Stars are of many types, some of them far more luminous than the Sun. Rigel [A], type B8, for example, a massive luminous star at the upper left of the main sequence, is 49,000 times as powerful as the Sun, the supergiant Antares [B], of type M, in the giant branch, has a diameter of 420 million km (260 million miles) and its luminosity is 3,400 times that of the Sun. Aldebaran [C], type K, 90 times more luminous than the Sun, is a less extreme red giant, with a diameter of 67 million km.

3 The evolution of a polar-type star is shown in this series of graphs. The star contracts out of the interstellar matter [A]. It then joins the main sequence [B]. After perhaps 10,000 million years it leaves the main sequence and moves into the giant branch [C], increasing in luminosity to 1,500 times that of the Sun and expanding in diameter to 50 times that of the Sun. It then becomes unstable and matter is ejected [D]. Subsequently the star collapses into a small, extremely dense white dwarf [E]. The red line follows the star's evolutionary track.



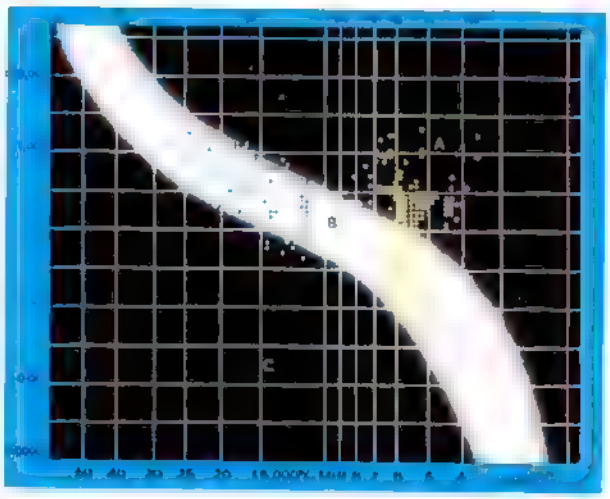
100,000 times that of water. After another long period all light and heat leaves the star and it becomes a dead black dwarf.

There is no positive information about black dwarfs, which send out no detectable radiation and we can only hazard guesses about their numbers. White dwarfs, however, are known to be common. The first to be identified (by Walter Adams [1876–1956] in 1916) was the companion of Sirius, which had been discovered by Alvan Clark (1832–97) more than half a century earlier, but had always been assumed to be cool and red. The surface temperature of the companion is greater than that of the Sun, but its diameter is only three times that of the Earth so that an immense amount of matter – almost as much as is in the Sun – is packed into a relatively small globe. Other white dwarfs since found are even more dense.

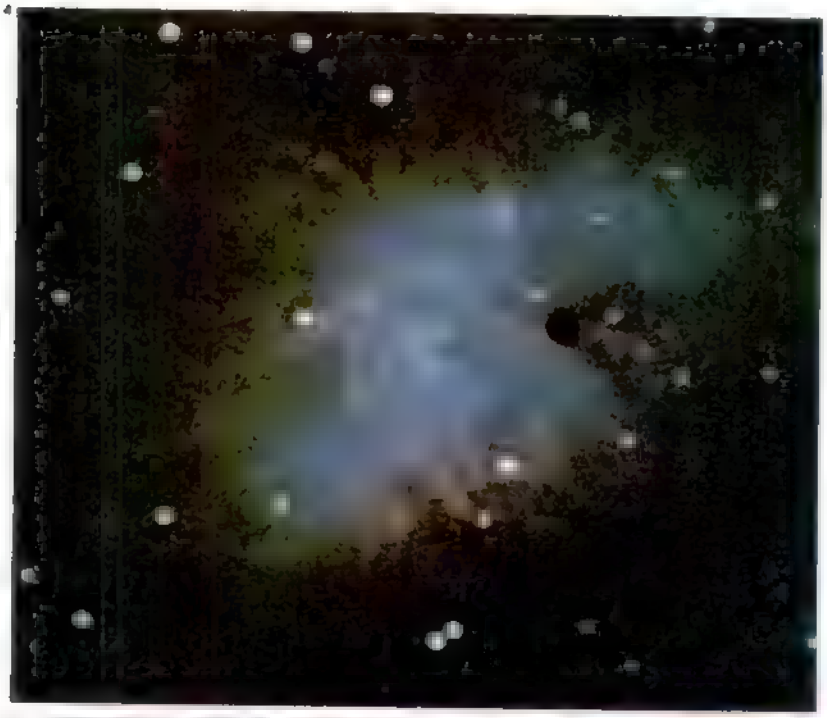
Evolution of a massive star

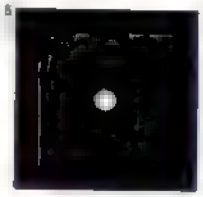
A star with a mass much greater than the Sun's [6] evolves much more rapidly. The luminous S Doradûs in the Large Magellanic Cloud, for example, cannot go on pouring forth energy at its present rate for much longer than a million years, whereas the Sun will not leave the main sequence for at least another 5,000 million years and stars of lower mass change even more slowly.

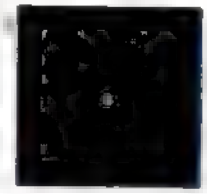
Very massive stars do not merely collapse into white dwarfs. When the core temperature has reached about 5,000 million degrees centigrade there is a catastrophic change in structure; the core collapses and the outer layers of the star, in which nuclear reactions are still going on, are abruptly heated to about 300 million degrees centigrade. The result is a supernova outburst in which the star emits as much energy in a few seconds as the Sun does in millions of years. Material is ejected and when the convulsions are over all that remains is a cloud of expanding gas together with a neutron star or pulsar, even smaller and denser than a white dwarf. The Crab Nebula is a supernova remnant; the outburst was watched by Chinese observers in 1054. Of the two nebulae shown here, the Rosette [1] represents a stellar birthplace while the Crab [4] shows the death of a once glorious star.

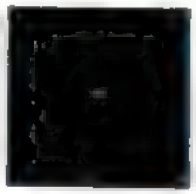


This Hertzsprung–Russell diagram in simplified form shows the main classes into which stars are grouped during their evolution. Typical stars in the giant class (A) have evolved out of the main sequence (B). At the lower left (C) are white dwarfs. Luminosity is measured vertically on the diagram and spectral type across the top.





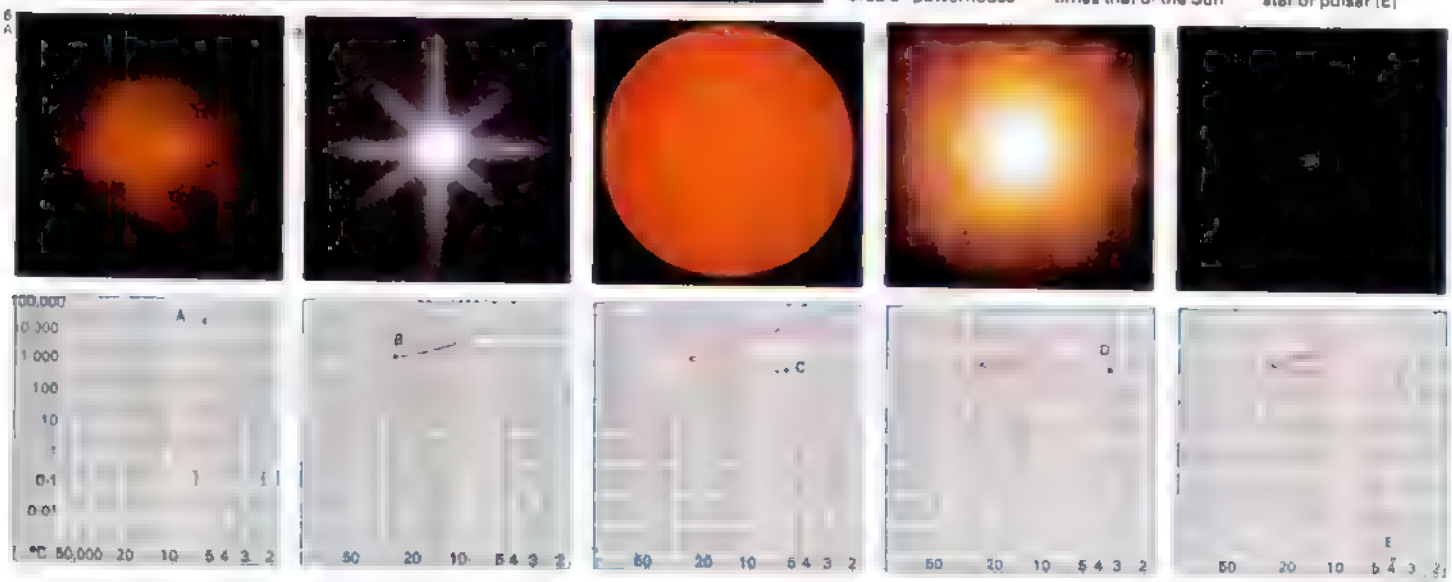




4 The Crab Nebula in Taurus, number one in Messier's catalogue, is an exceptional object. The supernova of 1054 became so bright that it could be seen in daylight with the naked eye, but after it faded below the sixth magnitude it was inevitably lost. The association between the 1054 supernova and the modern Crab Nebula has been questioned but seems to be no longer in doubt. The Crab contains the only pulsar to have been identified with an optical object, the pulsar has been termed the Crab's "powerhouse".

5 Stars on the main sequence, such as the Sun, are classed as dwarfs to distinguish them from members of the giant branch. The Sun (A) is a typical G-type main sequence dwarf. Capella is also of type G, but on the giant branch, with a luminosity 150 times that of the Sun. It is quite different from a white dwarf such as the companion of Sirius (B) which has collapsed, presumably from the giant stage, or a red dwarf such as Wolf 359 (C), one of the feeblest stars known, with a luminosity 0.00002 times that of the Sun.

6 The evolution of a massive star that is one with an initial mass of more than three times that of the Sun is shown. The star contracts out of the interstellar material (A) and joins the main sequence (B). After a period that is much shorter than for a solar-type star it moves into the giant region of the H–R diagram (C), "burning" first helium and then heavier elements. Eventually it experiences a supernova explosion (D) and sends most of its material outwards, leaving a neutron star or pulsar (E).



Galactic nebulae

Nebulae (clouds) are of various kinds and have proved to be of the utmost importance in modern astronomical theory. They appear in several parts of the sky as luminous patches that look like shining mist

Catalogues of nebulae

Several catalogues of nebulae have been drawn up by astronomers over the years. One of the most famous is that published by the French astronomer Charles Messier (1730–1817) in 1781. It contains more than 100 objects. Ironically, Messier was not in the least interested in nebulae, he was a comet-hunter who compiled his catalogue of nebular objects in order to avoid confusing them with possible new comets. In the late nineteenth century an extensive catalogue based on the observations of William Herschel (1738–1822) and his son John (1792–1871), was compiled by the Danish astronomer Johan Dreyer (1852–1926). It is known as the New General Catalogue (NGC). Today the NGC numbers and Messier's numbers (M) are still used

Messier catalogued all the nebulous objects, from star clusters to gaseous nebulae and to the systems such as the Andromeda Spiral, M31, that we now know to be galaxies. Astronomers have now agreed that the term "nebula" should be confined exclusively to clouds of gas or dust to avoid confusion

Galactic nebulae are of two main kinds emission and reflection. Both types occur not only in our own Galaxy, but in others. The so-called Tarantula Nebula lies in the Large Cloud of Magellan – 30 Doradûs (listed by Dreyer as NGC 2070) – and is much larger than the Orion Nebula, M42 [Key, 8], which is the most famous nebula in our Galaxy. The main constituent of all nebulae is hydrogen, which is the most abundant substance in the entire universe, but there is also a great deal of dust and it is this that absorbs starlight. Inside some of the nebulae are objects that cannot be seen, but can be detected by infra-red photography. Becklin's Object in the Orion Nebula is an excellent example. It may well be a star of tremendous luminosity, but it is permanently concealed from view

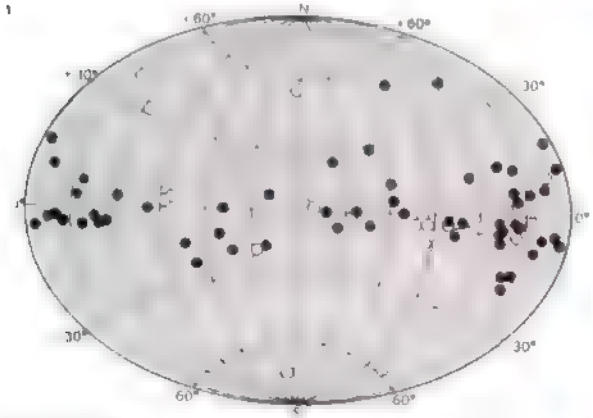
Vast though they are, the nebulae are made up of extremely tenuous material. The gas is many million times less dense than the air we breathe. It has been calculated that if a 2.5cm (1in) core sample could be taken right through the Orion Nebula the total weight of material collected would be no more than that of one small coin

Luminosity of nebulae

A nebula depends for its luminosity upon the presence of stars that are either close to it or are contained in it. If the stars are extremely hot, the hydrogen in the nebula is ionized and emits a certain amount of light of its own [Key, 2, 3, 8]. (Certain spectral lines in nebulae were once thought to indicate the presence of an unknown element, but it was subsequently found that the lines are due to familiar elements, such as oxygen, produced under unfamiliar conditions.) If the stars are less hot, the nebula shines only by reflection [5, 6]. If there are no suitable stars, the nebula does not shine at all, it remains dark and can be detected only because it blots out the light of stars beyond [5, 7]. There are various galactic nebulae within the range of small telescopes, although the vivid colours

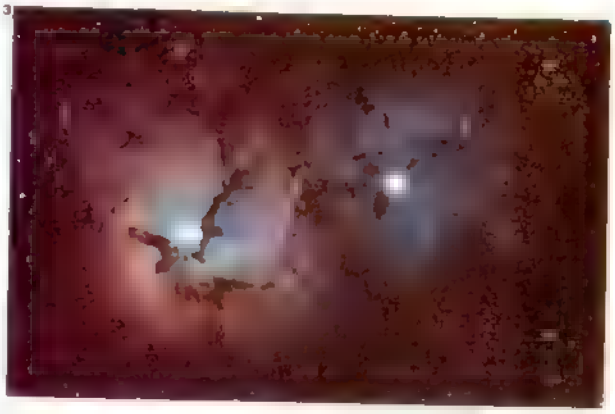
CONNECTIONS

See also



1 Gaseous nebulae are not distributed equally all over the sky. They occur in Population I regions in galaxies these regions are located in the main plane or spiral arms of the galaxy. Population II regions occur in the nucleus of the galaxy. Seen from the Earth the galactic plane is identified by the Milky Way – the area of the sky where most nebulae are found. The diagram shows the Milky Way region (galactic latitude 0 degrees) with the nebulae indicated by circles – black for reflection nebulae and white for emission ones. Only a few nebulae are shown, but the distributions are clear

2 An emission nebula emits light characteristic of the substance of which it is made up, if the nebulosity is illuminated by a suitably placed star. A star of the type W, O or B can cause ionization of the hydrogen atoms out to about 500 light-years, which is known as an H II region. If the temperatures of the associated stars are too low there will be no emission from the nebula



3 A famous emission nebula is the Trifid in Sagittarius, M20 (NGC 6514), shown in this Palomar photograph. Visible in a small telescope, it is 30 light-years across and more than 3,000 light-years away

4 The main galactic nebula in Sagittarius is M8 (NGC 6523), known as the Lagoon Nebula. Described by John Flamsteed (1646–1719) in 1680, it is easy to view telescopically because the integrated magnitude is 6.0. M8 is a dense nebula with 10^3–10^4 atoms per cubic centimetre in the central region. It lies some 4,850 light years from Earth. Associated with it is the galactic star cluster NGC 6530. M8 contains a number of T Tauri variables and also some dark globules that may eventually start to shine. Each globule has a diameter of about one light-year. Flare stars in M8 are also known. Moreover, M8 is a source of radio emission

of the photographs below cannot be seen visually. The colours themselves are genuine, but are too faint to impress the eye.

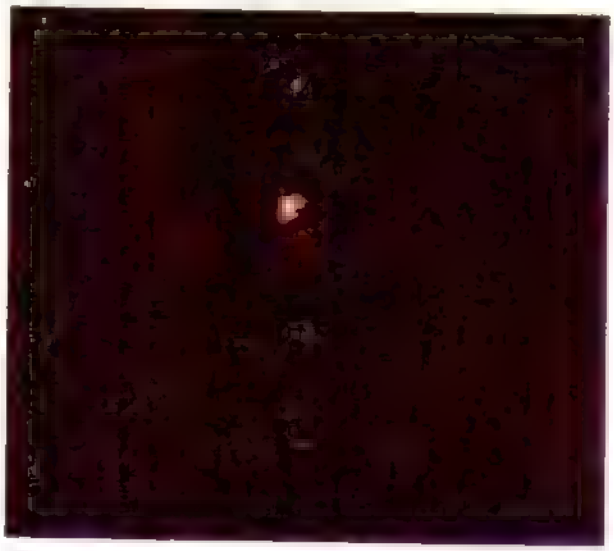
According to a classification produced by Walter Baade (1893–1960), there are two kinds of regions in our Galaxy (and in other galaxies): he called these Populations I and II. In Population I areas [1] there is a great deal of interstellar material and the brightest stars are hot and white. In Population II areas the interstellar material has largely been used up in star formation and the brightest stars are red giants. Because these stars are well advanced in their evolutionary careers, Population II regions seem to be relatively old. Gaseous nebulae occur in Population I areas, so their stars are presumably young by cosmic standards.

The formation of stars

The most important feature of Population I areas from the theoretical point of view is that they are apparently regions in which star formation is in progress. According to current ideas, a star begins its career by condensing out of interstellar material. Nebulae are obvious sites for such activity because the material in other regions of space between the stars is much too tenuous. On average interstellar space contains one atom of matter per cubic centimetre; nebulae, although rarefied, are more condensed than this. Objects such as the Orion Nebula, the Lagoon Nebula [4] and the Trifid Nebula [3] are in fact stellar birthplaces. The same is true of galactic nebulae in other systems, such as the Large Cloud of Magellan and the nebulae observable in the Andromeda Spiral. Dark patches in nebulae, known as globules, may well be embryo stars.

Nebulae also contain many stars that are variable in light and are unstable; these are known as T Tauri variables and are thought to be stars at an early stage in their careers that are still contracting towards the main sequence. Some stars have even been seen to increase in luminosity over a period of years, presumably because they have blown away their original dust clouds. One of these is FU Orionis, in the Orion Nebula, which became brighter in 1936 and must be one of the youngest stars known to us.

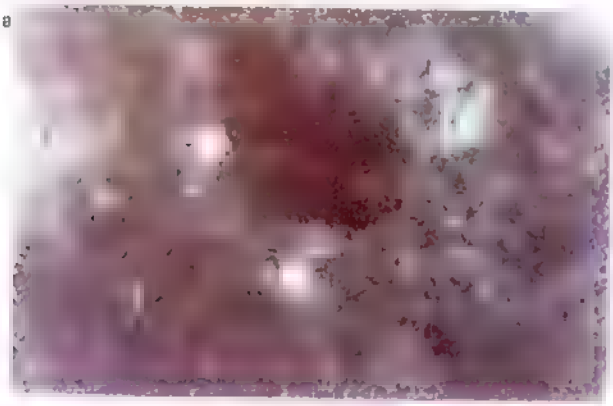
KEY



The Sword of Orion [M42] is the most spectacular of the gaseous nebulae. Within the nebula the famous multiple star, the Trapezium, responsible for the nebula's luminosity is clearly visible. The brightest star is Theta Orionis.

5 A

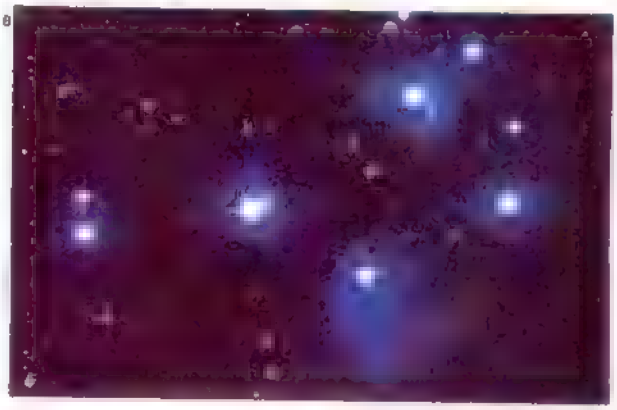
B



5 A dark nebula [A] cuts out the light of stars that lie at a greater distance from Earth. The Coal-sack [B] in the Southern Cross is the best example. The light of more distant stars is completely obscured, because light is absorbed by the nebula's solid particles, not by interstellar gas.

6 A

B

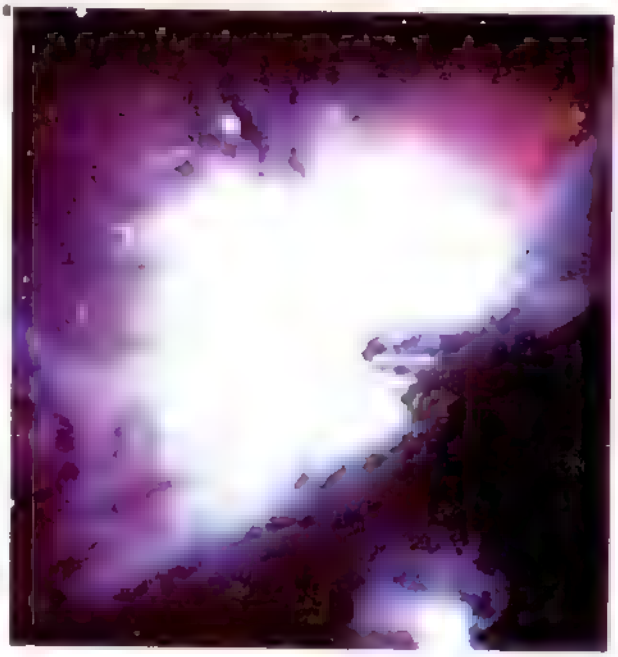


6 A bright nebula [A] shines by reflecting the light of a suitable star. The Pleiades cluster [B] in Taurus is an example of a reflection nebula. Such nebulae usually have a high dust content. The nebulosity in the Pleiades is best examined by using long-exposure photographs.

7

7 The Horse's Head Nebula in Orion lies close to Zeta Orionis, the most southerly star of the Belt. This dark nebula, shaped rather like the head of a knight in chess, is No 2024 in the New General Catalogue. It can be seen with a small telescope although it is best studied in photographs such as this Palomar picture.

8



8 The bright Orion Nebula to the south of the Belt in the 'Hunter's Sword' can be seen with the naked eye. Its luminosity is due chiefly to the multiple star Theta Orionis, on the side of the nebula turned towards the Earth. If this star did not exist, the nebula would be just as dark as the Horse's Head Nebula.

From nebulae to pulsars

Until fairly recent times it was not generally realized how different the various types of nebulae are. Looking at the Omega Nebula in Sagittarius and then the Crab Nebula in Taurus, for instance, leads to the conclusion that they are much the same; but in fact they represent diametrically opposite ends of the stellar evolutionary sequence. In Omega – a diffuse nebula – stars are being forced out of the interstellar material; the Crab represents supernova explosion debris at the centre of which is a pulsar – a collapsed dense star

Nebulae: the early stages

One interesting nebula is associated with the young star T Tauri, which is irregularly variable and is still contracting towards the main sequence The nebula was discovered in 1852 by J. R Hind, an English amateur who was using a 7in (17 8cm) refractor to hunt for asteroids and merely noted the T Tauri nebulosity in passing. Nine years later it was found that the nebula had disappeared. It has since been seen again and is within range of large telescopes (its official designation is NGC 1554), but it is not nearly as prominent as it was when Hind discovered it. Moreover T Tauri itself is not hot enough to excite the nebular material to self-luminosity; it is, however, an infra-red source and there is no doubt that it is associated with the nebulous material around it from which it has been formed. There are other variable nebulae similarly associated with young stars; for instance, R Monocerotis in the Orion area and R Coronae Australis in the southern sky

These, then, are nebulae involved in stellar birth. So are the familiar gaseous or galactic nebulae, such as M42 in Orion. Deep inside M42, permanently hidden from Earth by the nebular material, is an infra-red source known as Becklin's Object. It may be either a very young star or else an extremely powerful object at least a million times as luminous as the Sun; there is, however, no way of knowing because it is possible to study only its infra-red radiation, which can pass through the nebula and reach the Earth. In any case, star formation is in progress in the Orion cloud

A. Blaauw and W. W. Morgan, in America, have studied an O-type star, AE Aurigae, which has the high velocity of about 130km (80 miles) per second It looks faint only because it is remote; it is in fact a luminous young star. Tracing its path "backwards", so to speak, indicates that about two and a half million years ago it was in the region of the Orion Nebula [1] Moving in the opposite direction is another O-type star, Mu Columbae, which has similar velocity and is about equally far from the nebula. It has been suggested that a colossal disturbance probably hurled these two stars violently outwards from their place of birth

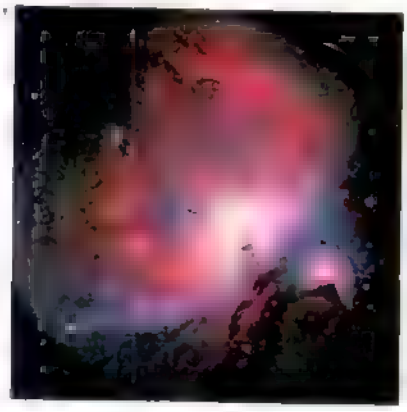
Planetary nebulae

Other nebulae represent later stages in the evolution of a star; in particular there are the planetaries, which look like small, feebly luminous discs or rings, not unlike those of planets. The planetaries, like the diffuse nebulae, are gaseous, but they are neither planets nor nebulae so their popular name could hardly be less apt The best known is the Ring Nebula, M57 Lyrae [4], which was discovered in 1779. It is made up of a central star surrounded by a spherical gaseous shell

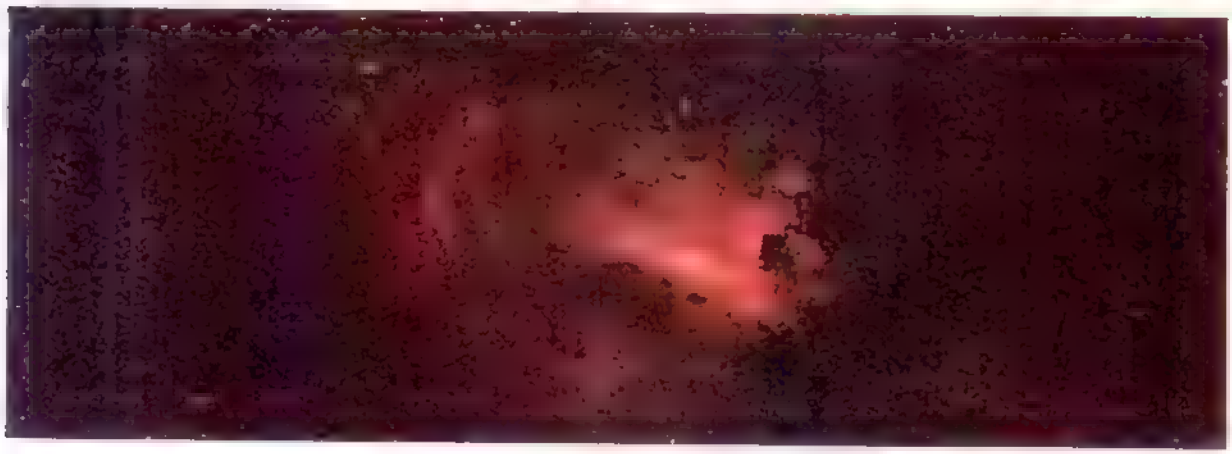
CONNECTIONS

See also

1 The Orion Nebula, a stellar birthplace, is the most famous of all gaseous nebulae This photograph was taken with the 200in (508cm) Hale reflector at Palomar USA. The "hollow" in the border to the right-hand side is due to the presence of the multiple star Theta Orionis, which lies close to the Earth-turned edge of the nebula. There is no doubt that fresh stars are condensing out of the nebula



2 The gaseous nebula M16 (NGC 6611) lies near the boundary between Sagittarius and the small constellation of Scutum, which adjoins the "tail" of Aquila M16 has an integrated magnitude of 6 4, so that it is visible with binoculars the distance from Earth is 5,900 light-years Seen through a foreground of stars the nebula, photographed in red light, shows both bright and dark nebulosity



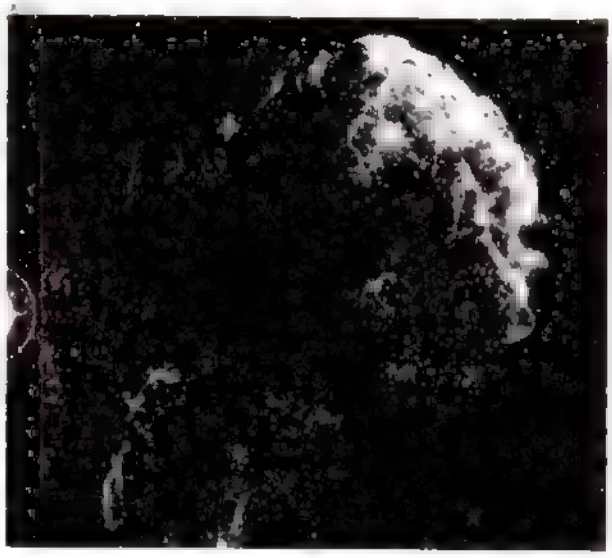
3 The Omega Nebula M17 (NGC 6618), sometimes known as the Horseshoe Nebula, was discovered by the French astronomer L. de Chéseaux in 1746 It is an easy binocular object, 1 5°N and 2°E of the 5th-magnitude star Gamma Scuti. It is more massive than the Orion Nebula, like so many other diffuse nebulae it has bright areas as well as signs of dark obscuring material M17 lies on the borders of the constellations of Sagittarius and Scutum.

that is incredibly tenuous. Looking at it, more is seen of the glowing material at the edge than at the centre so that the nebula looks like a ring. The diameter of M57 is almost one light-year but the gaseous surround is immensely rarefied, millions of times less dense than the Earth's air at sea-level. Some planetary nebulae are larger; for instance NGC 7293 Aquani [5] is twice the size of M57. Other planetaries are asymmetrical; these include the Owl Nebula, M97 Ursae Majoris, and the Dumbbell Nebula, M27 in Vulpecula.

All planetaries are expanding and their age can hardly be more than a few tens of thousands of years, it has been estimated that if the gaseous shell is ejected from an old star – as is likely to happen – the material cannot continue to shine for more than 100,000 years or so. According to one theory, a planetary is produced by a red giant star "puffing off" its outer layers, so that the central stars in planetary nebulae represent the cores of old giants. These stars have high surface temperatures of about 50,000°C and have completed their main nuclear burning; they are well on the way to becoming white dwarfs. The "puffing out" theory fits in well with the proposed evolutionary sequence, although it is by no means certain that every normal star inevitably becomes a planetary nebula at a late stage in its life.

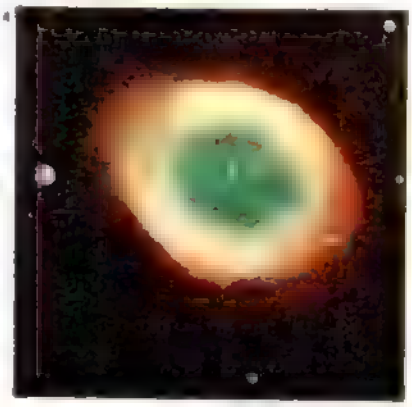
Supernovae and pulsars

Finally there are nebulae that represent the end products of stellar evolution. Although the Crab Nebula is the best known example, there are others but almost all are much older than the Crab and so their forms are not as well marked. (In any case, the Crab, with its unusual central pulsar, seems to be an exceptional object.) With the Veil Nebula [6], in Cygnus, the arched shape of the luminous material is plain and all the evidence points to the conclusion that it is the debris of a supernova outburst that took place in prehistoric times. The present rate of expansion is 120km (75 miles) per second. There is thus a full sequence of nebulae, from those of the T Tauri type associated with stellar birth through to the stellar remnants of supernovae explosions.

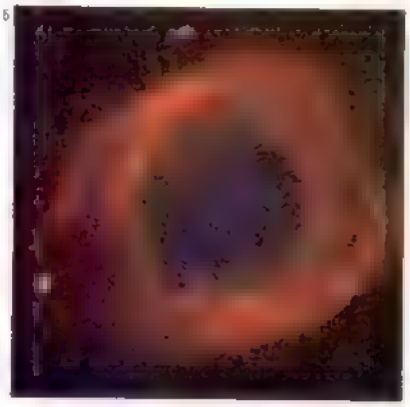


This gaseous nebula in Gemini is probably associated with the death of a star, because it gives every impression of being the remnant of a supernova outburst. Other types of nebulae are the birthplace of stars.

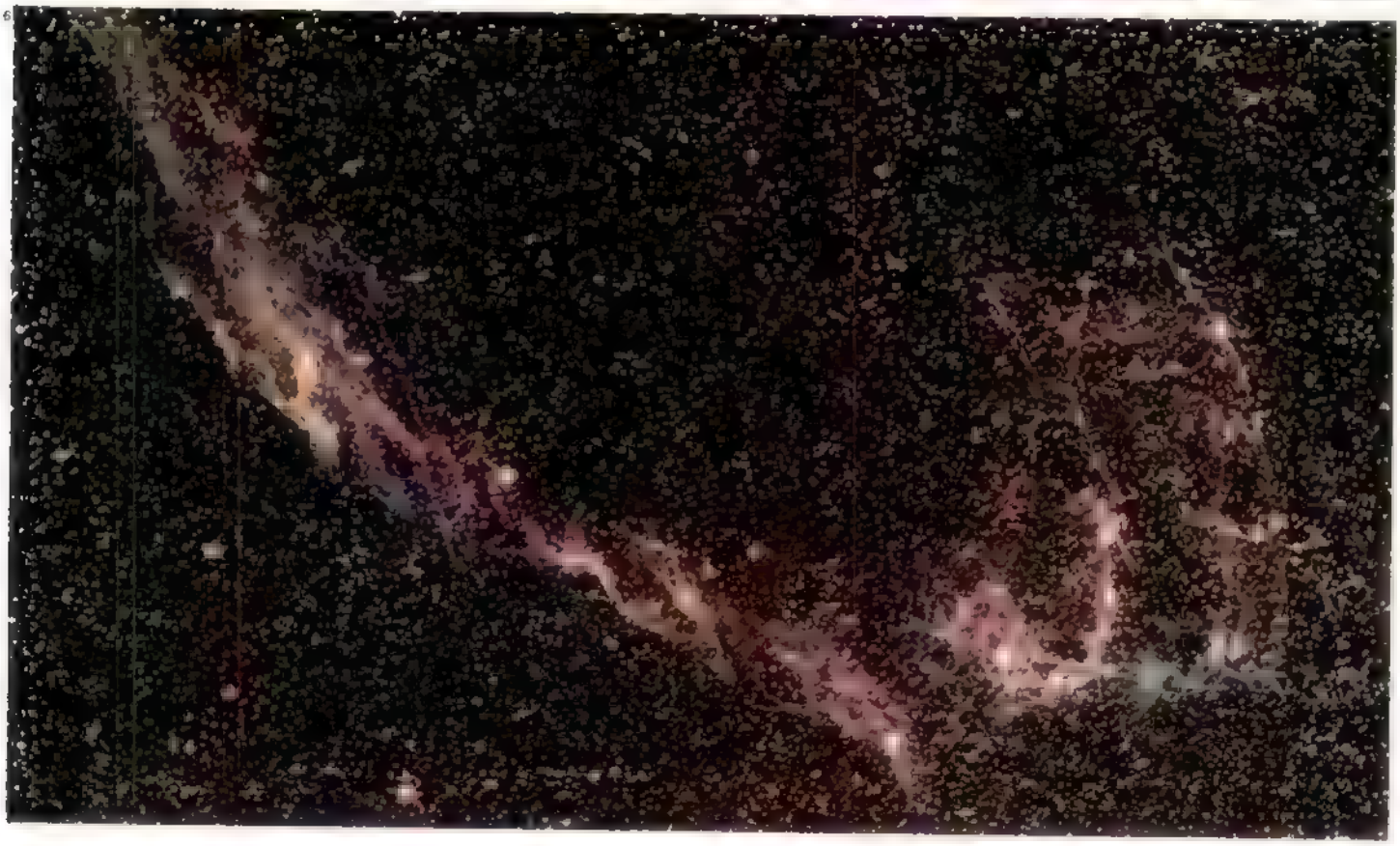
4 **M57 (NGC 6720), the Ring Nebula in Lyra** – the most famous of the planetaries – lies midway between the two naked-eye stars Beta and Gamma Lyrae. The variation in colours is due to temperature differences. The integrated magnitude is 9.3 and the distance from Earth 1,400 light years. The central star, although clearly shown, is by no means bright; the second star (upper right) is in the foreground and not connected with M57.



5 The planetary nebula NGC 7293, in Aquarius, was surprisingly left out of the famous catalogue of nebula objects drawn up by Messier, yet it is actually the brightest of all planetary nebulae and is definitely superior to the Ring Nebula in Lyra. It is 600 light years from Earth. The central star is of magnitude 13.3 and is clearly shown in this photograph taken with the 48in (122cm) telescope at Palomar.



6 The Veil Nebula in Cygnus, NGC 6992 – sometimes called the Cirrus Nebula – may be a supernova remnant. This photograph (48in (122cm), Schmidt telescope, Palomar) shows the arched shape, which is significant. The nebula is 2,500 light years away and from the motions in the gas it has been calculated that the supernova outburst occurred 50,000 years ago. In 25,000 years time the nebula will cease to be luminous.



Pulsars and black holes

To explain the nature of a black hole, it is necessary to examine stellar evolution [1]. A star such as the Sun contracts towards the main sequence. When the core temperature has risen sufficiently, nuclear processes begin. After the supply of available "fuel" has run low, the star swells to become a red giant, after which it collapses to the white dwarf condition.

A more massive star will behave differently and when its nuclear reserves fail it will explode as a supernova, ending its luminous career as a neutron star, or pulsar, together with a cloud of expanding gas.

From white dwarf to black hole

In a white dwarf the atoms are crushed, broken and packed together so tightly that there is little waste space. In a neutron star the gravitational field is so intense that the protons and electrons are forced to combine with each other to form neutrons, the density of neutron star material far exceeds that of a white dwarf. There is now little doubt that the radio sources known as pulsars are really neutron stars. The pulsar in the Crab Nebula has been identified with an optical object [Key, 3], and in 1977 the Australians identified another pulsar – in the constellation of Vela – with an extremely faint object of magnitude 26.5.

If a massive star collapses, it can pass through densities that correspond to the white dwarf and neutron star stages and still continue to contract, becoming steadily smaller and denser and entering a state of gravitational collapse where no known physical process can halt the contraction. Light will find it more difficult to escape and soon the body will contract within a critical radius (known as the Schwarzschild radius) at which point its gravitational field will become so strong that not even light will be able to move away from it. The star is then surrounded by what might be termed a "forbidden zone" from which nothing can escape. This is a black hole – a region that acts as a centre of gravitational attraction.

Inside a black hole, all the normal laws of physics break down. It has been suggested that the collapsed star may eventually be crushed out of existence altogether. And there have been sensational predictions that black holes may extend until they swallow everything in the universe. But ideas of this kind are, at best, highly speculative.

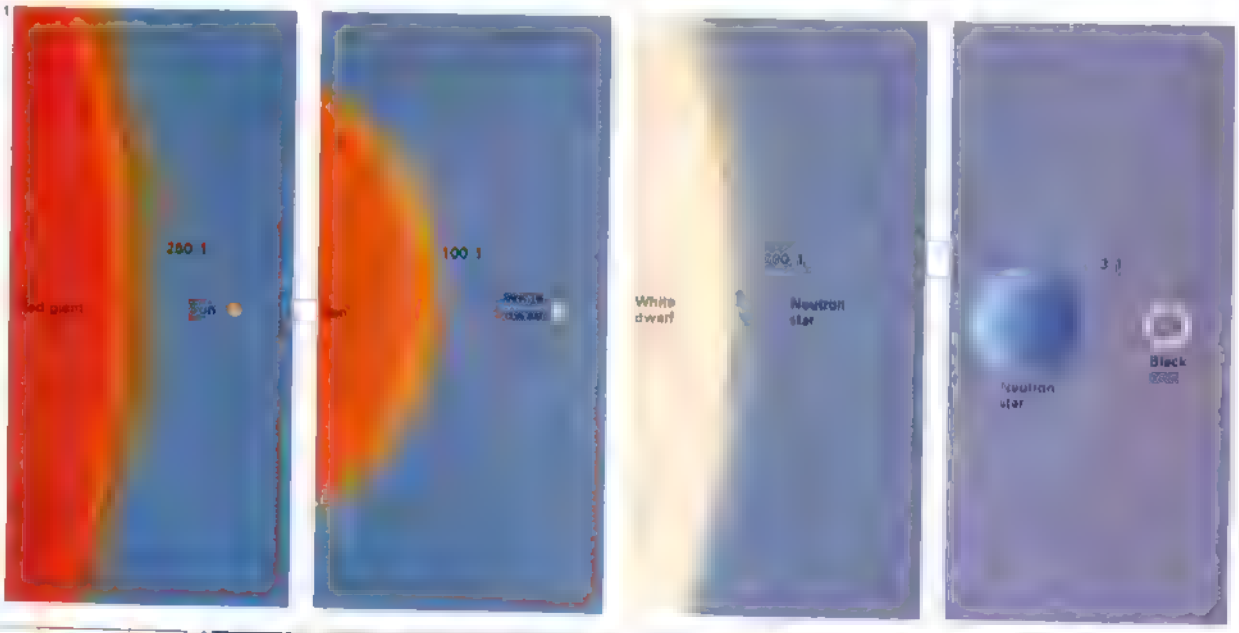
In search of black holes

The most promising place to look for a black hole is in a binary system. Close beside the brilliant yellow star Capella is a small triangle of stars, known popularly as the Haedi or Kids. At the apex of the triangle is Epsilon Aurigae, which is always clearly visible to the naked eye although it is by no means brilliant. In 1821 it was discovered to be variable, with a magnitude range of 3.3 to 4.2. Later it was found that Epsilon Aurigae is an eclipsing binary of unusual type, for the eclipses take place only once in 27 years and last for more than 700 days.

The brighter member of the pair is a highly luminous yellow supergiant, 60,000 times as powerful as the Sun. The fainter component, which causes the eclipses, has never been seen; it radiates only in the infrared and until recently all astronomers thought that it must be a large, cool star, still

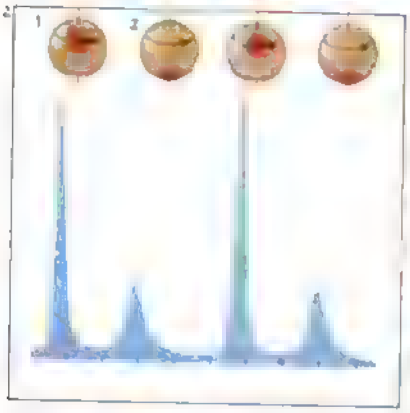
CONNECTIONS

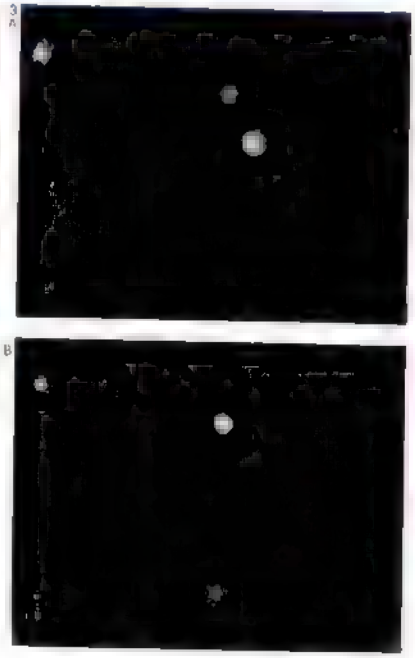
See also



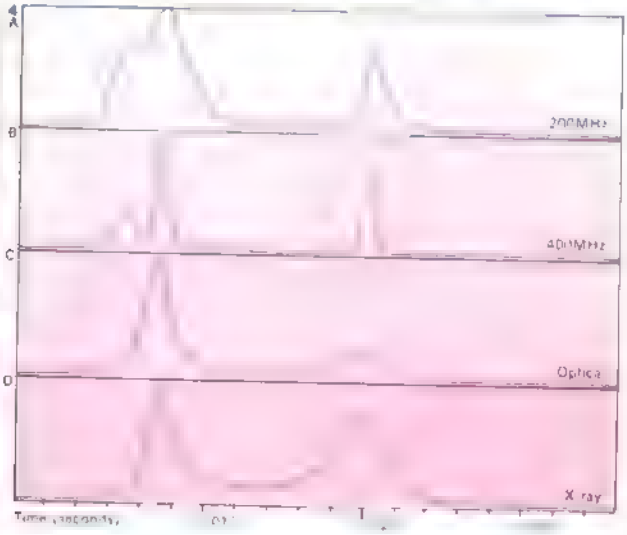
1 The relative sizes of a giant, the Sun, a white dwarf, a neutron star and a black hole are shown. The ratios are given for each diagram, thus the diameter of the Sun is approximately 100 times that of a white dwarf. The neutron star has the same mass as that of the Sun.

2 A pulsar's radiation varies. The red on the diagram indicates the ends of the magnetic axis of a pulsar. As the pulsar rotates, the signal strength varies according to the position of the axis. When one end faces Earth [1], the intensity is at its maximum. When the other end faces Earth [2], intensity is at its minimum.





3 The Crab Nebula pulsar is of great importance to astronomers, for so far it is one of only two to have been identified with an optical instrument. These photographs were taken at the Lick Observatory, California, with its 120in (305cm) reflector. The pulsar is easily identifiable in A but in B it is almost completely invisible. The whole pulse cycle amounts to only 33 milli seconds. There is now little doubt that pulsars are, in fact, neutron stars.



4 Radiation at all wavelengths is emitted by the Crab pulsar. The pulse cycle at a wavelength of 200MHz in the radio range is shown [A] and at 400 MHz [B], in the optical range [C] and in the X-ray range [D]. The beam of radiation has similar characteristics over the whole electromagnetic spectrum.

shrinking after its condensation from interstellar material, but not hot enough to shine by nuclear power. Now, however, there are suggestions that the infra-red member of the Epsilon Aurigae system may be a black hole.

This infra-red component seems to have a mass 2.3 times that of the Sun, which by stellar standards is high. It should therefore be luminous – but it is not. According to the American astronomers A. G. W. Cameron and R. Stothers it is a black hole, surrounded by a cloud of solid particles that are spiralling round the critical boundary – or "event horizon" as it is termed [7] – and are sending out the infra-red radiation detected from Earth. In time the particles will cross the event horizon and will enter the black hole, from which they never emerge.

X-ray sources

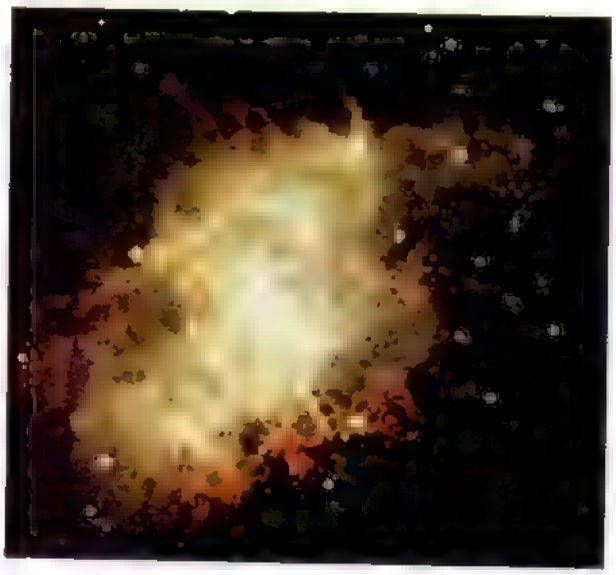
Another possible black hole is the companion of a supergiant star in Cygnus, known by its catalogue number, HDE 226868. The companion is a source of X-rays and it has been suggested that these are produced by material falling in towards the black hole and being accelerated to extremely high velocities.

X-ray astronomy is a recent development because it involves sending equipment above the shielding layers of atmosphere. The technique dates back to the 1960s. Many X-ray sources have been found, one of which is the Crab Nebula [Key]. Apparently most galactic X-ray sources are members of binary systems and are neutron stars associated with giants. There are also what are termed X-ray novae, which flare up, last for weeks or months, and then fade away.

Most of the X-ray sources are members of our Galaxy and lie reasonably near the main plane of the Milky Way, but some other galaxies also emit X-rays, notably the massive system in Virgo known as Messier 87. It is also a source of radio emissions.

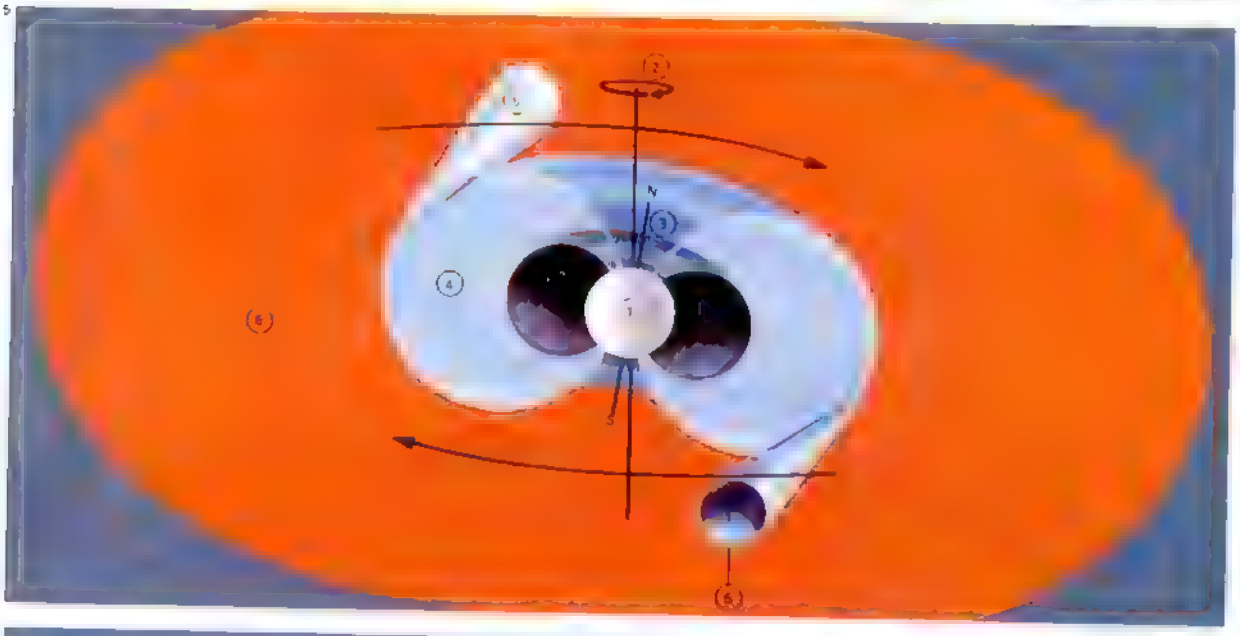
Progress in astronomy has been amazingly rapid during the past few years. In 1960 quasars and pulsars had not been detected and black holes were only of theoretical interest. X-ray studies had scarcely begun and even radio astronomy was primitive by modern standards.

Key



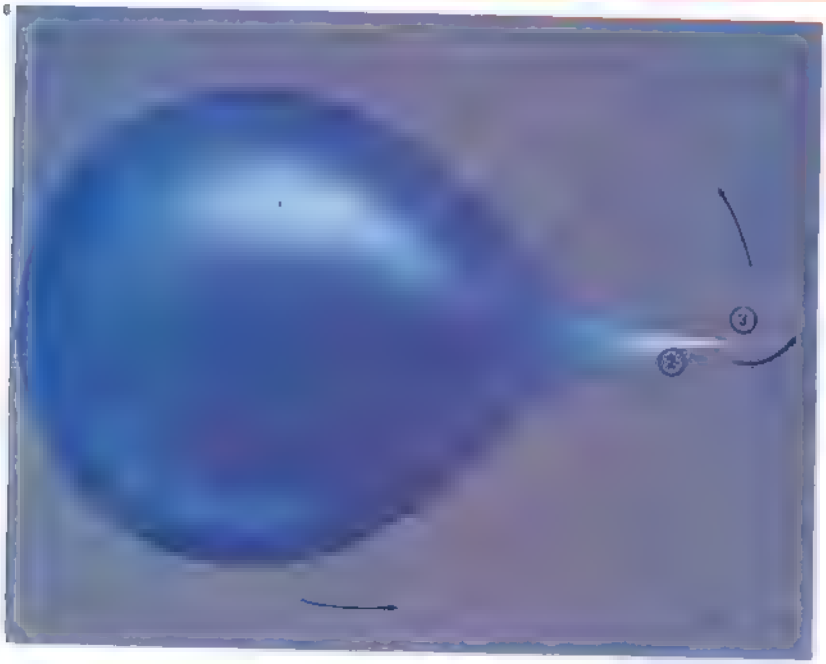
The Crab Nebula (M1, NGC 1952) is the remnant of a supernova in 1054. It contains a pulsar said to be the "power house" of the Crab, which is 6 000 light years distant.

5



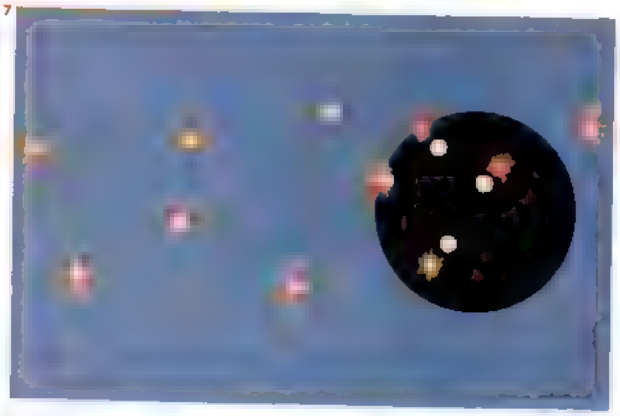
5 When pulsars were discovered, it was thought that the signals might come from rotating white dwarf stars. Now it is certain that a pulsar is a rotating neutron star [1], whose axis of rotation [2] does not coincide with the magnetic axis [3]. As the star the plasma rotates [4], sending out radio waves in beams [5]. Beyond this the plasma is stationary [6]. It is now agreed that it is the magnetic field of the rotating neutron star that generates the pulses as it turns over and over. The mechanism is related to the region some distance from the neutron star where the magnetic field would have to travel at the speed of light to keep up with the rotation.

6



6 The blue supergiant [1] it is believed has material pulled away from it in a jet [2] by the black hole [3], which is so small it can be regarded as a point mass. Because of its small size, the black hole takes in material only slowly. Before this, the material is compressed and subject to great heat, so emitting powerful X-rays. The invisible companion of HDE 226868 in Cygnus may be a black hole with a diameter of about 100km (60 miles). The supergiant has a diameter 300 000 times larger.

7



7 The "event horizon" is the boundary of a black hole (black area). The origins of some light sources are shown with locations of each light a moment later (white circles), depicting how gravity effectively bends light as it falls into the black hole.

Double stars

Our Solar System is centred on a single star, the Sun, but the universe contains many stars that appear to be close pairs or even members of complicated systems.

These double stars are surprisingly common, but are not always exactly what they seem. Some are indeed binary, or physically associated systems, the association of others is an illusion, the result of a mere line-of-sight effect. If two stars, as seen from Earth, happen to lie in much the same direction, they will appear to be side by side in the sky, even though there is no real connection between them. An example is Vega, the brilliant blue star in Lyra. It has a twelfth magnitude companion that is much farther away, although, from a terrestrial point of view, it is apparently close.

Binary stars and their structure

It was originally thought that all double stars must be the result of a line-of-sight effect. Not until 1793, and the observations of William Herschel (1738–1822), were true binary pairs discovered. In a binary system, the two components move around their common centre of gravity. For some pairs, the period of revolution is short – in extreme cases, less than 20 minutes – while for others it is long.

Gamma Virginis, not far from Spica, is made up of two exactly equal components with a revolution period of 180 years. The angular separation is now less than it was earlier in this century because the two stars are moving closer to the same line of sight. The pair used to be separable with any small telescope, but by AD 2016, when the apparent distance between the two will be at its minimum, Gamma Virginis will appear to be single, except in giant telescopes.

Mizar (Key), and its companion Alcor, in Ursa Major, is a particularly easy binary to spot – it was the first double star discovered by telescope. Like Alpha Centauri, it has two rather unequal components, one of magnitude 2.4, the other of 3.9.

Some pairs, such as Gamma Arietis, have the same spectral type for both components, but others are distinguished by their beautifully contrasting colours. Antares, the brilliant red star in the Scorpion, has a faint green companion and the same is true of the red giant star Alpha Herculis [3, 4]. But perhaps the best example is Beta Cygni [8, 9], or Albireo, which has a golden-yellow primary and a greenish-blue companion.

Spectroscopic and eclipsing binaries

If the separation between the components is slight, the binary will appear single. However, the revolution of the two components round their common centre will show up in the spectroscope [10]. The brighter component of the Mizar pair is a spectroscopic binary system.

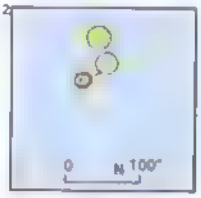
There are also systems made up of more than two stars. Alpha Centauri [1, 2], for instance, the closest of all the bright stars, is made up of two rather unequal components (of magnitudes 0.0 and 1.7) and has a revolution period of 80 years. Closely associated with it is Proxima Centauri, making Alpha into a triple star. Proxima is the nearest star to Earth, but it is much less bright than Alpha. Epsilon Lyrae, near Vega, is an example of a wide pair, each component of which is again double. Castor, in Gemini, is a sixfold system, in which four of the components are bright

CONNECTIONS

See also

1 Alpha and Beta Centauri are the pointers to the Southern Cross, but are too far south to be seen from Europe

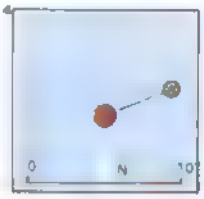
2 Alpha Centauri is a triple star, visible through a telescope.

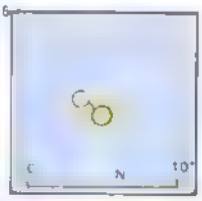


3 Hercules is not a brilliant constellation. However it contains three interesting features: the great globular cluster M13, just visible to the naked eye, and the two binaries, Rasalgethi and Zeta.

4 Alpha Herculis is a red giant (magnitude 3-4), its companion is greenish.

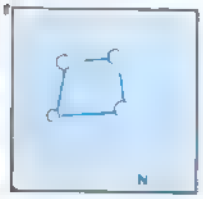
5 Zeta Herculis is another binary with unequal components (magnitudes 3.1, 5.6), its period is 34 years.

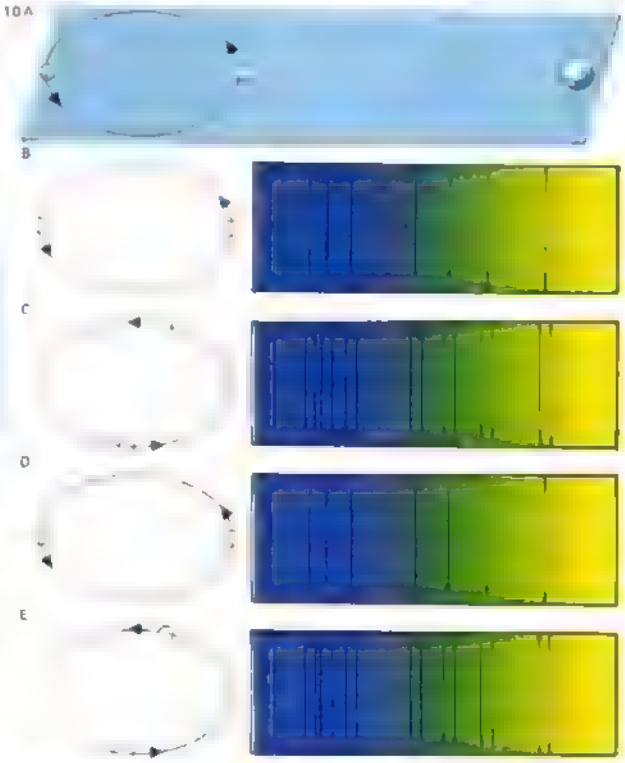




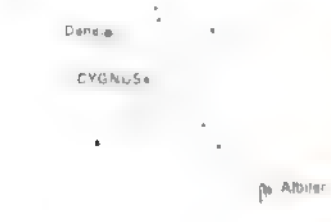
6 The pattern of Orion [A, B] shows two brilliant stars, Betelgeuse and Rigel. Rigel is one of the many double stars in Orion, which also contains the nebula M42, visible to the naked eye.

7 The multiple star Theta Orionis, nick-named the Trapezium, is contained in the outer part of the Great Nebula. All the components are of spectral type O and presumably had a common origin. They are visible with a small telescope.

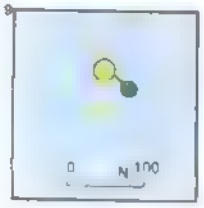




8 Cygnus is one of the richest of all constellations because it lies in the Milky Way. Its leading star, Deneb, is a highly luminous giant. The constellation takes the form of a cross, but the symmetry is spoiled by Albireo (Beta Cygni), a double with a yellow primary and also a greenish companion.



9 The golden-yellow primary of Beta Cygni is of type K, with an absolute magnitude of −2.2. Its companion is of magnitude 5.4. Despite their great separation these are still binary stars.



10 Analysis of a spectroscopic binary assumes that the stars are equal in mass, thus moving in circular orbit round their common centre of gravity [A]. The Earth, many light years away, lies in the plane of their orbit. The stars move transversely to the line of sight from Earth [B]. Then the lower star moves towards the Earth and its spectral lines shift to the blue (or violet), the upper star shows a red shift and moves away and the lines in the combined spectrum appear double [C]. The stars again move transversely and the lines merge [D]. The sets of lines again shift in opposite directions [E]. Periodic doubling of lines indicates a binary

and the other two are dim red dwarfs, it is made up of two spectroscopic binaries and a third, much fainter companion star that is also a binary.

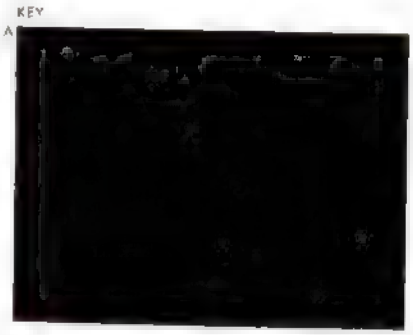
During the revolution of the two stars of a binary system, one component may pass behind the other, either totally or partially. When this happens, the light visible from Earth will be reduced and the star will seem to give a long, slow "wink". The prototype of these eclipsing binaries is Algol (Beta Persei) [11, 12], in which eclipses occur every 2.87 days and the magnitude drops from 2.2 to 3.5. Minimum magnitude lasts for 20 minutes and each fade – and recovery – takes five hours. Many stars of the Algol type are known. With Beta Lyrae [13, 14], near Vega, the components are close and less unequal. As a result there are two well-marked minima during the total period, which amounts to 12.9 days. Some eclipsing binaries have short periods - Delta Librae is, for example, is only 2.3 days. Others have long periods – 972 days for Zeta Aurigae near Capella, and as much as 27 years for Epsilon Aurigae in the same region [15, 16].

It is evident that there is no essential difference between an eclipsing binary and an ordinary system, everything depends on the angle from which observations are made. Seen from a different angle, Algol would not exhibit any eclipses at all and would appear constant in light.

The importance of binary stars

Binary pairs were formerly thought to be the result of the fission or breaking up of a single star that rotated so quickly as to become unstable. However, it is now thought more likely that the components of the binary stars are formed separately in the same region of space and at the same time.

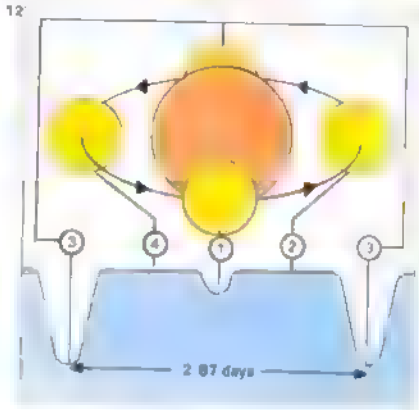
Binary stars make an important contribution to the total fund of astronomical knowledge. Measuring the mass of a single star is difficult, but observation of the components' orbital movements enables astronomers to estimate the combined mass of binary systems. Eclipsing binaries provide further opportunities for collecting data: study of their light-curves gives valuable insights into the diameters of the components.



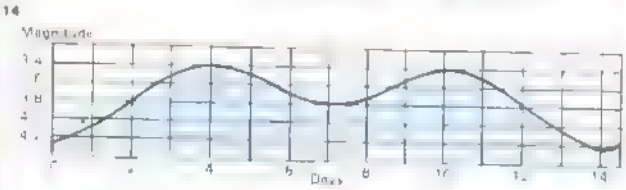
The 7 stars of Ursa Major [A] – the star pattern nicknamed the Plough – include Mizar, magnitude 2.4, or Zeta Ursae Majoris, in the Bear's tail. Naked-eye observation on a clear night will reveal Alcor, magnitude 3.9, apparently close beside Mizar. Through a telescope the binary system is seen clearly: there is an obvious difference between the primary and secondary components [B]. The system is further complicated – as the results of spectroscopic research show – for Mizar itself is another binary. Alcor is a true member of the system but it is so far from Mizar that orbital revolution takes millions of years.

11 Perseus contains some fairly bright stars and has a distinctive shape. Algol lies in the southern part of the constellation and has much fainter stars (Kappa and Rho) on either side of it.

12 The Algol binary system consists of a small, bright star of type B8 (yellow) and a large, fainter star of type K (orange). When as seen from the Earth the two stars shine separately [2, 4], a constant level of light is received. When the brighter star eclipses the fainter [1], the light received drops slightly. The least light is received when the fainter star eclipses the brighter [3].

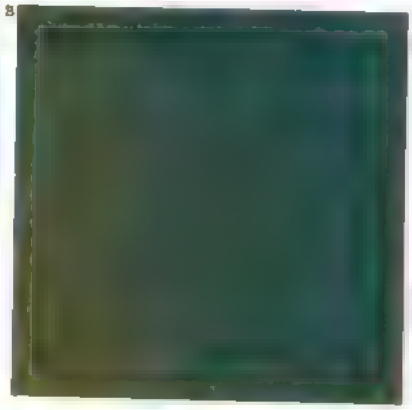


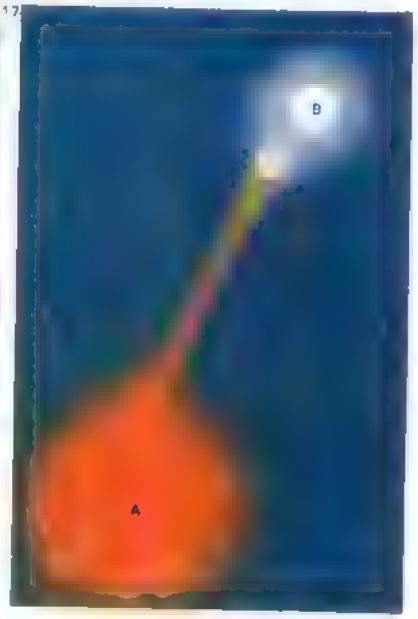
13 The brilliant blue star Vega, 26 light-years away, lies in the small constellation of Lyra. Fifty times as luminous as the Sun, Vega dominates the whole region. Close by is the quadruple star Epsilon Lyrae and a red, semi-variable R Lyrae. Beta Lyrae or Sheliak, the eclipsing binary, makes up a pair with its neighbour, the third-magnitude Gamma Lyrae.

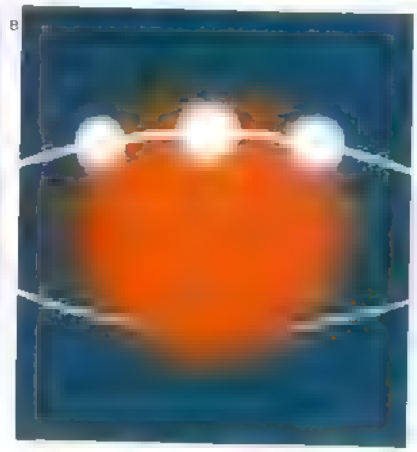


14 The components of Beta Lyrae are so close that they almost touch. Although they cannot be seen separately and are surrounded by complicated gas-clouds, it is known that they must be egg-shaped because of their nearness. Unlike Algol Beta Lyrae is always varying – there are two minima, one of magnitude 3.8 and the other 4.3, taking place alternately. The maximum magnitude is 3.4 and the period is 12.9 days.

15 The pattern of Auriga [A] is dominated by the bright yellow Capella. The two eclipsing binaries Epsilon and Zeta Aurigae lie in the small triangle close by Capella, but are not genuinely associated stars. Unlike many binaries, Epsilon and Zeta Aurigae are always visible to the naked eye [B]. The third member of the triangle, Eta Aurigae (magnitude 4), is a useful comparison star.







16 Zeta Aurigae is made up of an orange supergiant star (type K9, diameter 290 million km) and a hot white star (type B7, diameter 4 million km). When the bright star is eclipsed [A] (every 972 days), there is a three week period when it shines through the supergiant's tenuous outer layers, producing informative spectral changes. When the bright star passes in front of the supergiant [B] there is no marked decrease in light.

17 Dwarf nova binaries or U Geminorum stars are close binaries, in which one star is a member of the main sequence, the other a white dwarf. Material is pulled off the larger star [A] across to the white dwarf [B] in a jet, striking the gaseous surround and producing a spot [C] brighter than the stars themselves. Variations in this jet produce rapid flickerings in the light, detectable only with electronic equipment. Periodic outbursts also come from the white dwarf.

Pulsating stars

Pulsating stars are variables whose brightness wanes with time due to cycles of expansion and contraction. The variations may be regular or irregular, varying from a few minutes to many centuries. Astronomers have been continually engaged in monitoring and searching for new variables [2, 3]. A notable observer, John Goodricke (1764–86) was the first to realize that the curious "winking" behaviour of Algol, in Perseus, is due to the periodical eclipse of the bright star by a darker companion, and he discovered the variability of Delta Cephei, which has proved to be one of the really important members of the Galaxy, as far as theoretical research in astronomy is concerned.

Delta Cephei [5], in the far north of the sky, has a fairly small magnitude range of 3.6 to 4.3, so that it is never conspicuous and yet never becomes dim enough to be hard to see with the naked eye. Its period – the time between one maximum and the next – is 5.366 days and is absolutely regular, so that the brightness for any particular moment can always be predicted. Subsequently other stars of the same kind were discovered. Eta Aquilae in the Eagle, with a period of 7.17 days; Zeta Geminorum in the Twins, 10.2 days, and Kappa Pavonis [6B] in the southern constellation of the Peacock, 9.1 days. With modern methods many similar variables have been discovered – the known number now totals many thousands – and they have become classified as Cepheids.

The period-luminosity law

Cepheids are giant stars of high luminosity. Being well advanced in their careers they have become unstable. However they are quite unlike the explosive stars, whose behaviour cannot be predicted. But the Cepheids are of vital importance for one main reason: their changes in output provide a key to their real luminosities and hence to their distances.

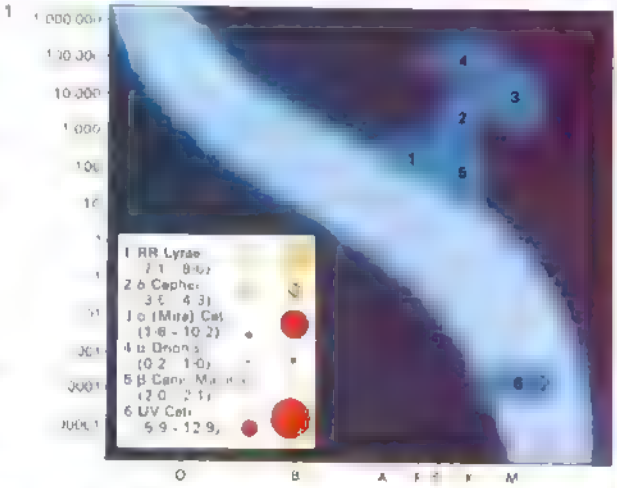
The key was discovered in 1912 by Henrietta Leavitt, who was working on some photographs of the external system known as the Small Cloud of Magellan [Key, 2]. The Cloud contains Cepheids, and Leavitt found that the stars of longer period looked brighter than those of shorter period. For all practical purposes the stars in the Cloud can be regarded as being equally distant from Earth, just as, two men in New York City, one standing in Times Square and the other by the Statue of Liberty, are equally distant from London or Paris – and so it followed that the brighter Cepheids were genuinely the more luminous. If a star's real power and apparent brightness are known, then its distance can be worked out. Naturally, many corrections had to be made (notably for the absorption of light in space), but the principle was clear and the Cepheid period-luminosity law has provided the main method of gauging distance in the Galaxy.

Beyond the Galaxy

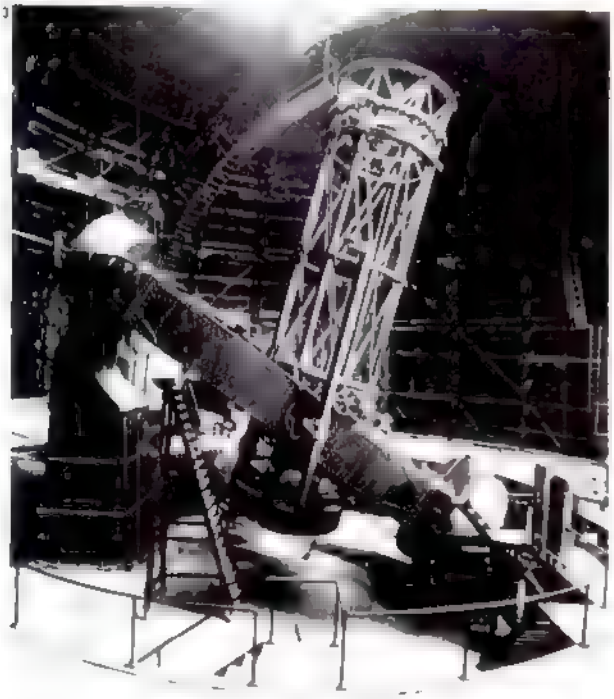
In 1923 Edwin Hubble (1889–1953) at Mount Wilson, found Cepheids in some of the "starry nebulae", including M31 in Andromeda. As soon as he had found their periods, he could obtain their distances. He realized that the Cepheids – and hence the spirals themselves – lay far beyond the limits of our Galaxy. Without Hubble's discovery of these convenient Cepheids, proof would

CONNECTIONS

See also

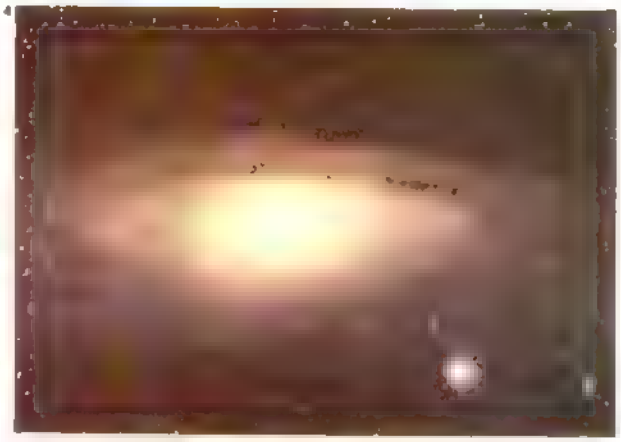


1 The Hertzsprung–Russell diagram shows a star's luminosity (or absolute magnitude) against its spectral type. Here the main classes of pulsating stars are plotted – RR Lyrae stars [1], Cepheids [2], the long period Mira stars [3], red variables [4], Beta Canis Minoris variables [5] and Flare stars [6]. The inset gives the variation in magnitude from minimum to maximum of the main members of each class. The Cepheids are of early spectral type, in a stage through which many stars are thought to pass.



2 The refracting telescope formerly in Peru and now at the Boyden Observatory in South Africa was used to obtain the photographs of the Small Cloud of Magellan from which in 1912 Henrietta Leavitt (1868–1921) established the period-luminosity law of the Cepheid variables.

3 The 100in (254cm) Hooker reflector at Mount Wilson, USA was completed in 1918. For 30 years the most powerful telescope in the world, it was used by Hubble in studies of short period variables in external galaxies; at that time, the 1920s, no other instrument was powerful enough for research of this kind. The 100in is still in full operation. The mounting is of the English type so that the telescope can never be pointed towards the celestial pole. The driving mechanism is powered by falling weights.



4 Trying to find short-period variables in external galaxies Hubble concentrated on the Andromeda Spiral – not then known to be beyond our Galaxy. As the most impressive of the spirals, it seemed likely to be relatively near by cosmic standards. Hubble failed to find RR Lyrae variables in it, but he was able to locate Cepheids. As soon as he had observed their periods he was able to use the period-luminosity law to show that the Cepheids – and hence the Spiral itself – must be beyond the Milky Way. His estimate was 750 000 light years but the real distance is 2.2 million light years.

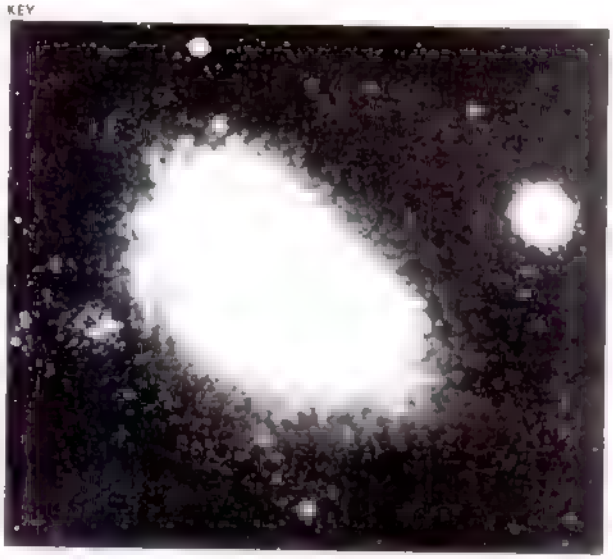
have been extremely difficult to obtain. It is true that his original estimates were found to be too low because of an error in the Cepheid scale that did not come to light until the work of Walter Baade (1893–1960) in 1952. Hubble believed the Andromeda Spiral to be 750,000 light-years away, but the real distance is more than 2 million light-years [4].

Because Cepheids are so powerful, they can be seen over immense distances and even at about 40 million light-years they are still detectable. There are also some associated variables of shorter period (less than a day) all of which appear to be of about the same luminosity, roughly 90 times that of the Sun. They are known as RR Lyrae stars, after the best-known member of the class.

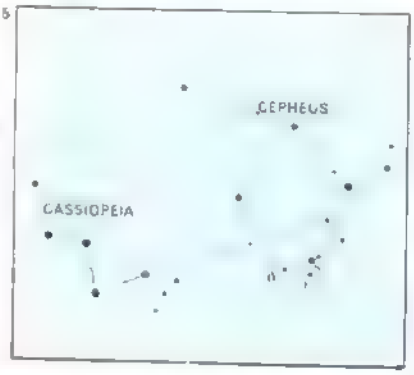
Long-period stars

Cepheids and RR Lyrae variables [7] are pulsating stars, alternately swelling and shrinking. There are also stars that pulsate in much longer periods, of from a few weeks up to a year or more. These are the long-period variables, often called Mira stars after Mira [10, 11, 12], the "Wonderful Star" in Cetus the Whale. Virtually all stars of this kind are old red giants of tremendous size and high luminosity; they have used up their available hydrogen "fuel" and are unstable. There is no Cepheid-type period-luminosity law and indeed the periods and the amplitudes are not constant. Mira itself has a period of 331 days, but this may vary by a week or so either way from one cycle to another. At some maxima Mira may become as bright as the Pole Star (magnitude 2), while at other maxima it is no brighter than the fourth magnitude. When at its faintest the magnitude is about 10, so that ordinary binoculars will not show it. Another Mira star is Chi Cygni, in the Swan, which ranges between magnitudes 3.3 and 14.2.

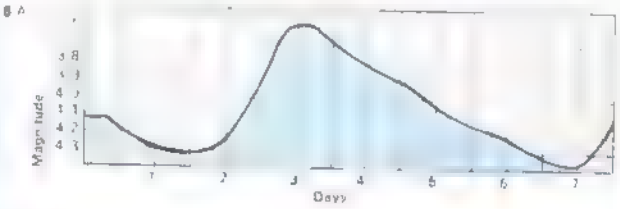
There are also semi-regular variables, such as Betelgeuse in Orion, with small amplitudes and periods that are very irregular indeed. Most, though not all, are red giants and they too swell and shrink, changing their output of energy as they do so. Stars such as Betelgeuse are extremely large, with diameters of up to 580 million kilometres (360 million miles).

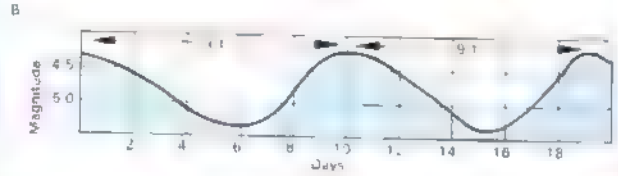


The Small Cloud of Magellan is in the southern sky, inaccessible from the great northern observatories. Photographs taken from Peru before World War I confirmed that it contains Cepheids, on which the period-luminosity law was based.

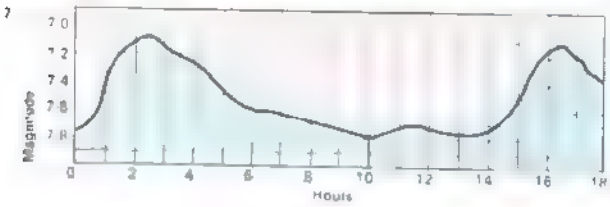


5 Delta Cephei lies in the far north of the sky. It forms a triangle with Epsilon and Zeta Cephei which act as convenient comparison stars; the fluctuations of Delta are obvious.

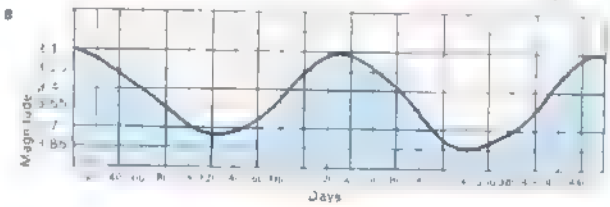




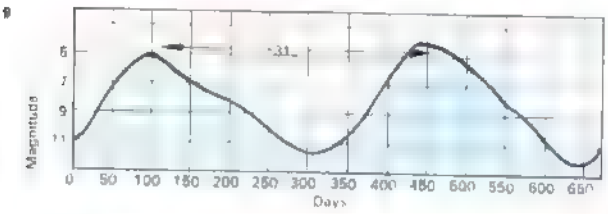
6 Cepheid variables are notable for their regularity. Delta Cephei itself has a period of 5.3 days [A], in comparison the light curve of the southern Cepheid variable Kappa Pavonis has a period of 9.1 days [B]. The shapes of the curves are not identical. Since Kappa Pavonis has the longer period, it is the more luminous. Other famous Cepheids are Zeta Geminorum (10.2 days) and Eta Aquilae (7.17 days).



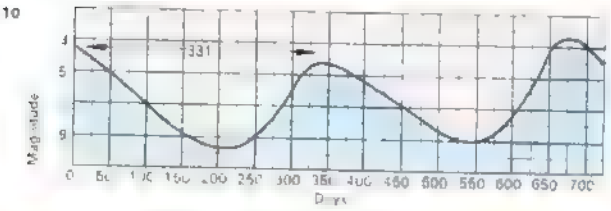
7 The periods of RR Lyrae stars, formerly called cluster Cepheids, are much shorter than those of the classical Cepheids; they are all of about the same luminosity.



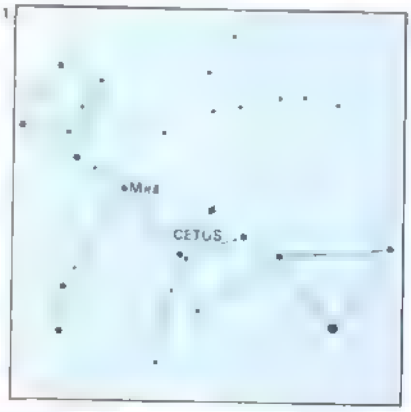
8 Eta Geminorum is a semi-regular variable with a small magnitude range; neither the period nor the amplitude is constant from one period to another but the fluctuations are relatively slight.



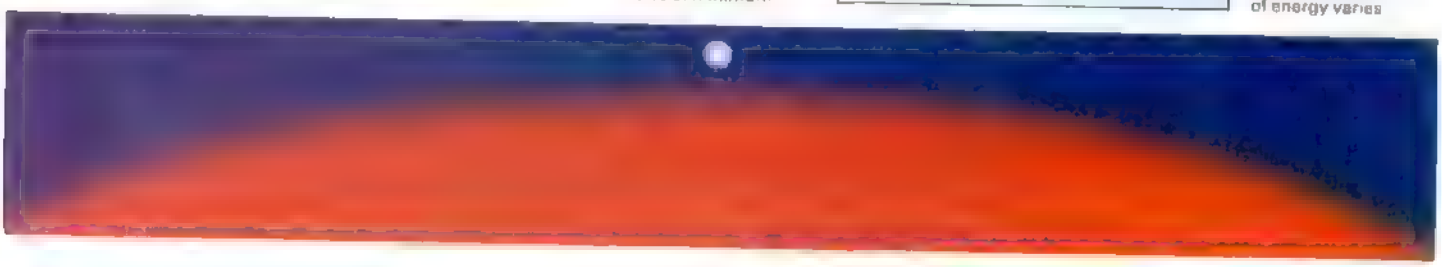
9 The light-curve of R Leonis is typical of long-period variables of the Mira type. As with all Mira stars both period and amplitude are subject to fluctuation. At its brightest, about magnitude 5, R Leonis is a naked-eye object.



10 Mira Ceti is another long-period variable that is visible to the naked eye. The average period is 331 days; the magnitude ranges from 1.7 to 4 at maximum and down to 10 at minimum.



11 Mira Ceti lies in a barren area of the sky and thus is easy to locate when at its brightest. But it is visible to the naked eye for only a few weeks each year although not regularly in each period.



12 The size of Mira is here compared with the Sun; it has a diameter like all red giants of more than 160 million km (100 million miles). The diameter changes as the star's output of energy varies.

Irregular stars

Not all variable stars are predictable. Like the regular pulsating stars, the many irregular variables are categorized into groups of similar type. For instance all semi-regular stars, such as Betelgeuse, have only approximate periods – the time elapsed between a star's magnitude at maximum and minimum. R Coronae stars [2] remain normally at maximum and suffer sudden, unpredictable drops to minimum. U Geminorum stars [3, 4] or "dwarf novae" are normally at minimum, but increase abruptly to maximum before fading away again. RV Tauri stars [5, 6] are G- to K-type giants and have alternate deep and shallow minima, superimposed upon which are periods of total irregularity. Flare stars [Key], for example the M-type giant UV Ceti, show sudden increases over periods of minutes and remain at maximum only briefly so that their changes can actually be watched. Recurrent novae show sudden, violent outbursts over periods of years, thus T Coronae exploded in 1866 and again in 1946. Normal novae [7–10] show one outburst only and then return to their former obscurity. The exceptional highly brilliant star Eta Carinae [11, 12] is classed as a pseudo-nova.

It is customary to draw light curves of irregular variables and novae in the same way as those for regular pulsating stars, relating apparent magnitude to time. (Apparent or visual magnitude is the apparent brightness of a celestial body as seen with the eye. The brighter the object, the smaller the numerical value of the magnitude.) It must be emphasized that the apparent magnitude of a star is its brilliancy as seen from Earth; it is not a reliable guide to a star's real luminosity. Only the variable stars and novae show short-term changes in apparent magnitude.

Semi-regulars and irregulars

Most of the semi-regular stars are red giants. They are regarded as unstable because they swell and shrink. Betelgeuse in Orion is one such star. Sometimes it will almost equal Rigel in brilliance, its mean magnitude (0.85) is comparable to that of Aldebaran. It has an approximate period of five to six years between maximum and minimum, but the irregularities are very marked. Rasalgethi, or Alpha Herculis, another semi-regular var able, is easily visible to the naked eye. Semi-regular stars that can be detected only by telescope are also common. Generally the variations in the magnitude are not great.

Most irregular variables are telescopic objects, however. Gamma Cassiopeiae [1] can rise to almost the brilliance of Castor in the Twins – as it did in 1936. The spectral changes were of great interest and apparently the star was throwing off a shell of material.

Probably the most erratic variable in the sky is Eta Carinae, in the southern hemisphere. During the middle of the nineteenth century it shone more brightly than any star in the sky apart from Sirius, but since 1867 it has been too faint to be seen with the naked eye, although binoculars will pick it up. It is orange-red and surrounded by nebulosity, when seen through a telescope it appears as a small patch rather than a sharp point like a normal star. It is luminous and remote.

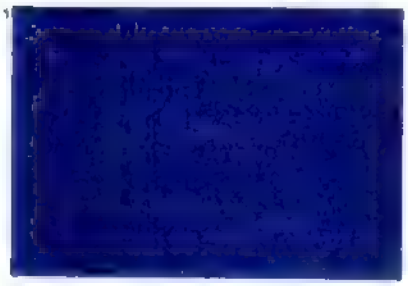
R Coronae and U Geminorum stars

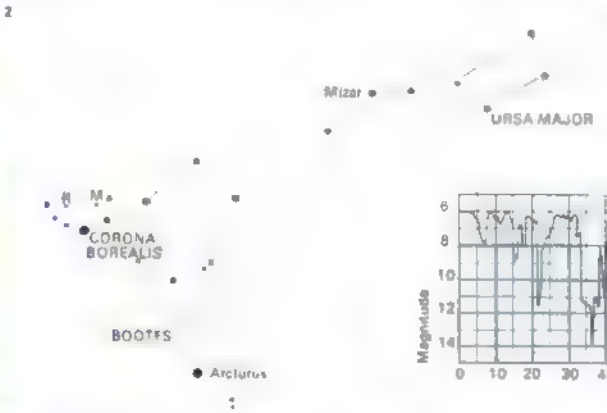
R Coronae Borealis [2], in the Northern Crown, is the prototype of one class, of which fewer than 50 members are known. R

CONNECTIONS

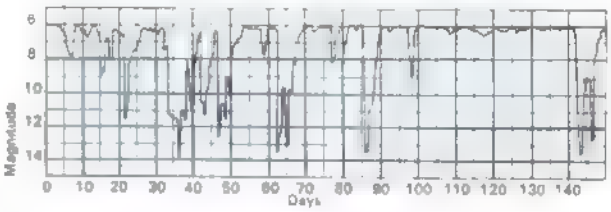
See also

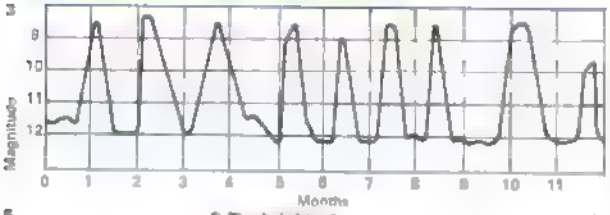
1 The W of Cassiopeia contains the irregular variable star Gamma Alpha Cassiopeiae or Shedir (spectral type K) is also suspected of slight variability, although with a small range of magnitude.



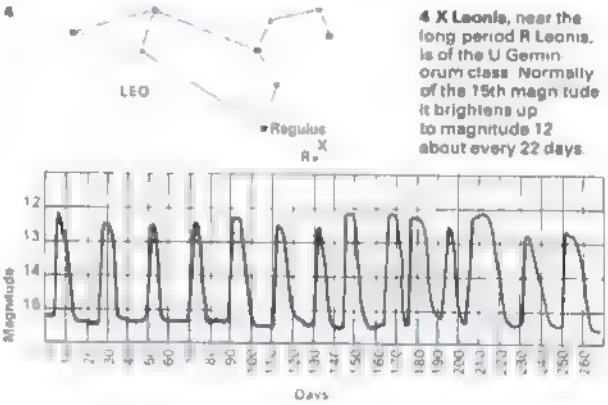


2 R Coronae lies in the bowl of the Northern Crown (Corona Borealis) not far from Arcturus. It is normally about the sixth magnitude, so that, together with the most useful comparison star (M) with a magnitude of 6.6, it can clearly be seen with binoculars. At some minima R Coronae drops to 15 and cannot be seen through moderate-sized telescopes.



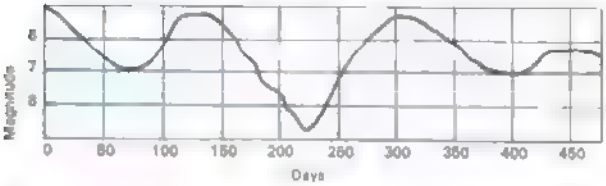


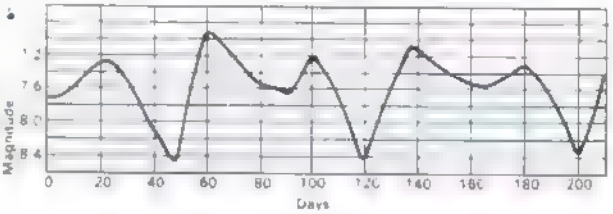
3 SS Cygni lies not far from the brilliant star Deneb. Out bursts occur on average every six weeks (from magnitude 12 to 8.25), but this may show wide fluctuation. SS Cygni is the brightest member of U Geminorum variables.



4 X Leonis, near the long period R Leonis, is of the U Geminorum class. Normally of the 15th magnitude it brightens up to magnitude 12 about every 22 days.

5 The brightest of the RV Tauri variables is R Scuti in the tiny constellation of Scutum. The Shield, near the tail of Aquila, the Eagle. It is easy to find because it is one of four stars making up a quadrilateral and is not far from the beautiful open cluster of M11, which is nicknamed the Wild Duck. The light curve shows magnitude plotted against time, but the curve is only an average because all RV Tauri stars are erratic in their behaviour. The range of R Scuti is between magnitudes 5 and 8.6 so that at maximum it is visible to the naked eye. It always stays within the range of binoculars, appearing reddish in colour. It is therefore a favourite object for amateur astronomers.





6 All the RV Tauri stars are highly luminous and are among the most massive variables known, some are at least 25 times as massive as the Sun. Unfortunately, all are remote and few are within binocular range. The alternate deep and shallow minima may sometimes be replaced by spell s of total irregularity. AC Herculis, whose light curve is shown here is a case in point, it has a magnitude range from 7 to 8.5.

Coronae stars stay at their maxima for most of the time but suffer sudden, unpredictable falls to minima. They are poor in hydrogen but rich in carbon, it has been suggested that the minima may be due to the accumulation of carbon particles in the star's outer atmosphere, which causes the radiation from the star itself to be temporarily shut in. R Coronae at maximum is on the fringe of naked-eye visibility.

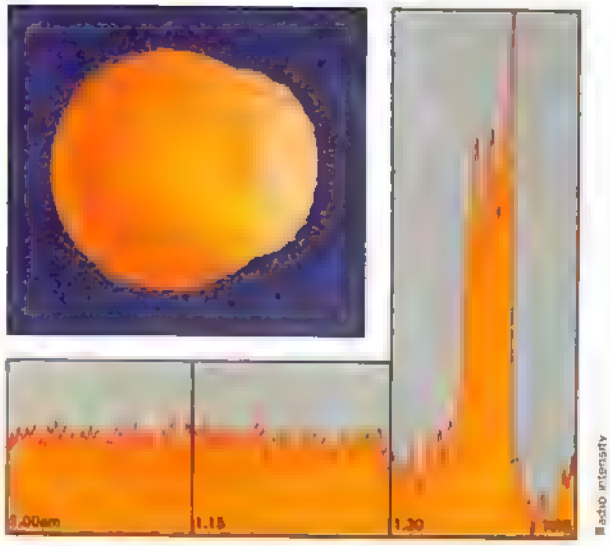
The U Geminorum or SS Cygni stars normally stay at minimum, but undergo periodical outbursts. The average interval between outbursts of SS Cygni [3] is about six weeks. It is now known that all SS Cygni or U Geminorum stars are close binaries with one white dwarf component together with a late type red dwarf.

Normal and recurrent novae

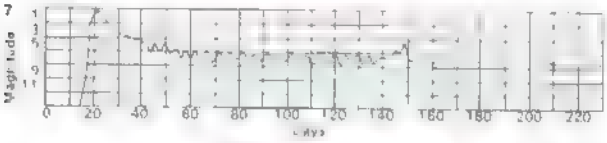
A nova is not a new star; it is a formerly obscure star that has suddenly increased in brilliancy. Some past novae have been brilliant, for instance both Nova Persei (1901) [7] and Nova Aquilae (1918) [8] exceeded the first magnitude at maximum. Once a nova has passed its peak it fades back to its original brightness, although it may take years to do so. It is thought that the outburst affects only the star's outer layers - whereas in a supernova explosion the star is destroyed in its old form. Many, perhaps all, novae are spectroscopic binaries.

HR Delphini [10], one of the most interesting novae of modern times, was discovered in July 1967 by the English amateur George Alcock. It never became brighter than magnitude 3.6, but it was slow to fade and remained a naked eye object for a year. By 1975 it had fallen to below magnitude 11, but it may not fade much more. It is one of the few novae whose pre-outburst magnitude of 22 is well known. Since it is about 30,000 light-years away, we are watching the results of an explosion that must have happened some 30,000 years ago.

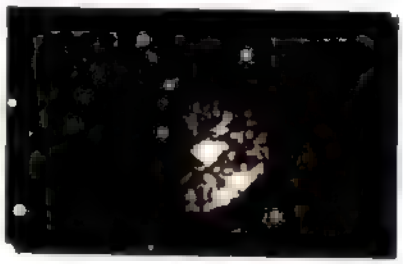
A few stars have been known to suffer more than one outburst. T Coronae blazed up from magnitude 9 to 2 in 1866 and from magnitude 10 to above 3 in 1946. Such stars are the recurrent novae, but not many of them are known.

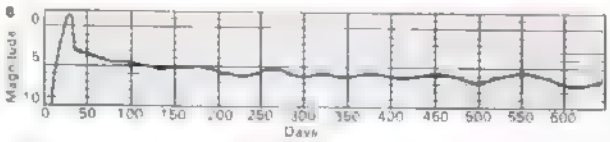


Flare stars, such as UV Ceti, are nearby dwarfs and, alone among variable stars, change in brilliancy so quickly that they can be watched as their luminosity increases over a period of minutes. The sudden increases are due to intense flare activity.

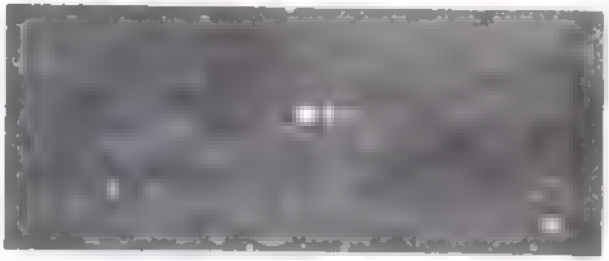


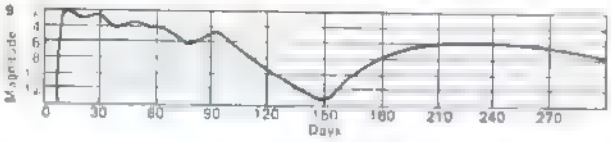
7 Nova Persei (1901) became very brilliant but declined rapidly. The decrease was attended by marked fluctuations. Nebulosity round the star was illuminated, giving the false impression of an expanding cloud. Today Nova Persei is a faint object.



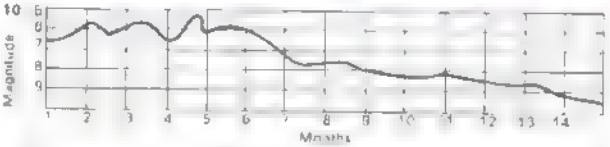


8 Nova Aquilae (1918) appeared with dramatic suddenness. At its maximum on 9 June, its magnitude was 1.4, equal to Sirius. It soon faded but remained visible to the naked eye until November 1918. A nebulous cloud round it steadily expanded and became dimmer, finally disappearing from view in 1940. The old nova is still visible, although extremely faint. Nova Aquilae was the brightest exploding star in modern times.



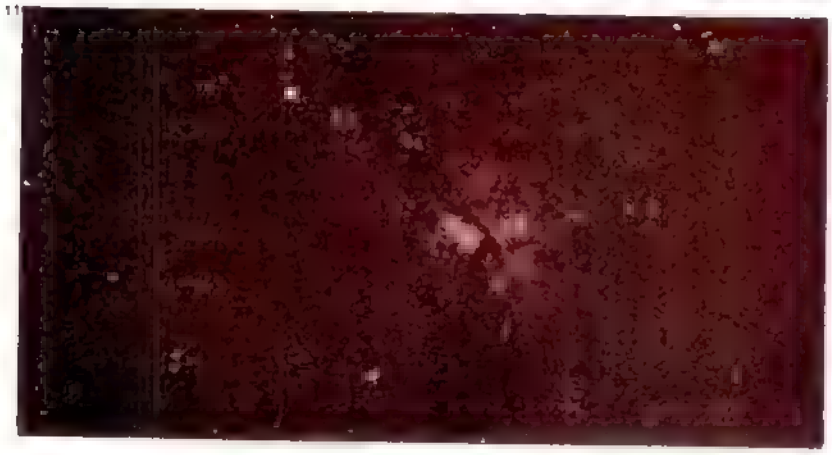


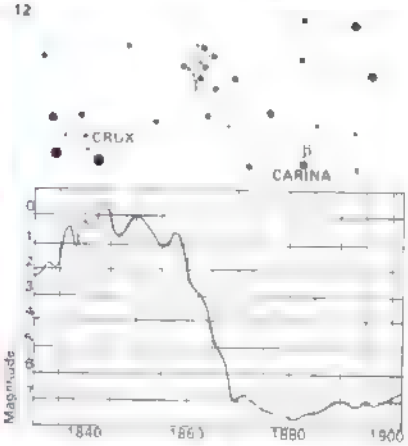
9 Nova (DQ) Herculis (1934) exceeded the 2nd magnitude and was unusual because it showed a temporary recovery after its drop from maximum. These photographs taken in 1951 show that the nova is a close binary with associated nebulosity. It is classed as a 'slow nova'.



10 HR Delphini (1967) has the distinction of being the slowest true nova on record. The maximum lasted for six months, although with definite fluctuations, and the subsequent fall was gradual. It is likely that it can be observed through a small telescope.

11 Eta Carinae is not a normal nova. It is intensely luminous and is associated with complex nebulosity. The fluctuations may be due in part to variations in density of the nebulosity, but the star itself must also be intrinsically variable. In the telescope it appears strongly red and seems to be much less well defined than a normal star.





12 In the Keel of the Ship, near the Southern Cross and the 2nd-magnitude star Beta Carinae, lies Eta Carinae. It is the most erratic of all variables. It reached its maximum brightness in the 1840s when it was exceeded only by Sirius. It then declined and since about 1867 has been invisible to the naked eye, although binoculars will show it. It is surrounded by nebulosity and is in every way an exceptional object of high luminosity.

Stellar clusters

Man lives in a part of the Galaxy in which the density distribution of the stars in space is about average. His nearest neighbour, Proxima Centauri, is more than four light-years away, and there are not many stars within a radius of ten light-years from the Sun. Here and there in the Galaxy, however, are groups of stars making up genuine clusters. The best known example is the cluster of the Pleiades or Seven Sisters [5], in Taurus; several others are easily visible to the naked eye, and the number of known clusters visible through a telescope is immense.

Open clusters

Clusters are essentially of two kinds: open and globular. The open or loose clusters are to be found in the spiral arms of our Galaxy and are irregular in shape; they may be rich with thousands of members, or poor with only a dozen or two. There is no possibility of their being caused by line-of-sight effects.

There are wide differences between the various open clusters. In the Pleiades cluster the brightest stars are hot and white, and there is a large reflection nebula indicating the presence of a great deal of interstellar material; by cosmic standards the group is extremely young. Several of its chief stars are known to be in rapid rotation and one of them, Pleione, is so unstable that it periodically sheds some of its material and produces a shell or gaseous ring. This can be studied only by spectroscopic methods, round its equator.

In the second of the Taurus clusters, the Hyades (round Aldebaran) [3], the star density is smaller, the principal members are not so energetic and the amount of material spread between the stars is less. The Hyades are not as spectacular as the Pleiades because they are largely overpowered by the brilliant orange light of Aldebaran. Yet Aldebaran is not a true member of the cluster at all and here a line-of-sight effect is found, because Aldebaran, in fact, lies midway between Hyades and ourselves.

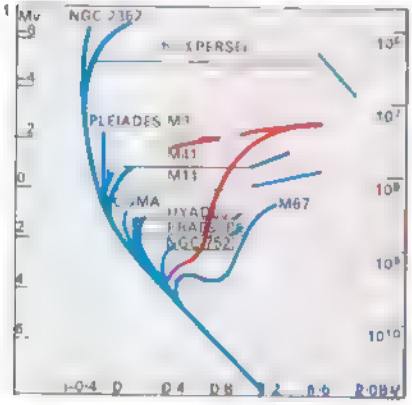
Other naked-eye clusters are Praesepe, the Beehive [4], in Cancer and the lovely southern cluster round Kappa Crucis, known as the Jewel Box because it contains stars of varied colours. In Perseus, not far from the W of Cassiopeia, is the Sword Handle [2], which is made up of two rich clusters in the same telescopic field.

Open clusters are not stable associations and must eventually be disrupted by the gravitational pulls of field stars in our Galaxy. It has been estimated that most of them have life-spans of no more than 1,000 million years before being scattered to the extent of losing their separate identity. One of the oldest clusters known, M67 in Cancer, which is easily visible with binoculars near the star Alpha Cancri, may be more than 4,000 million years old, but is located well away from the galactic plane and as a result is less liable to disruption.

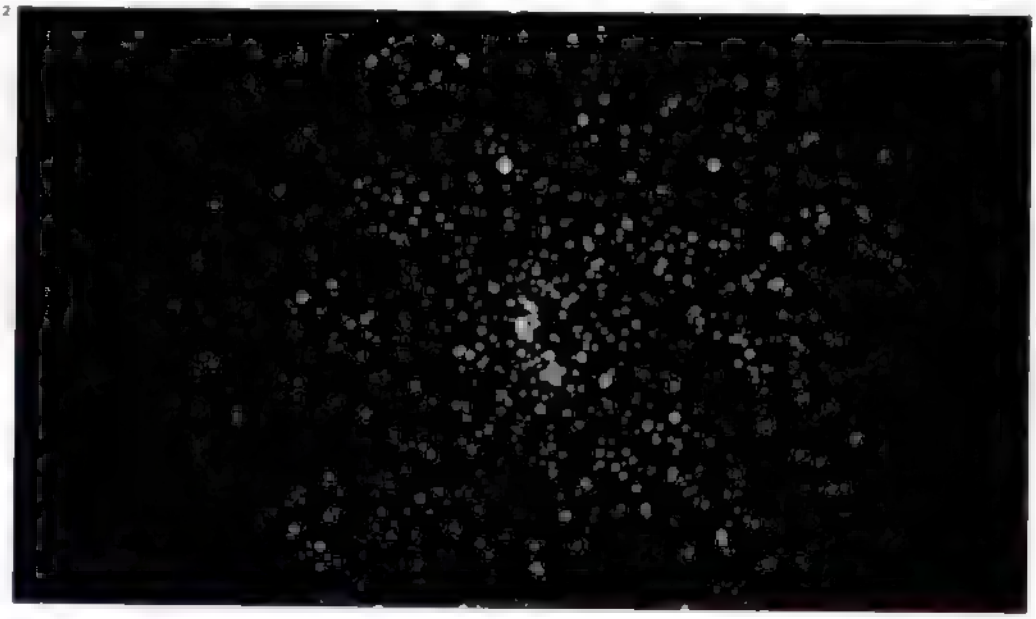
Globular clusters

Globular clusters are of a different type altogether. Only about 120 are known to exist in our Galaxy. They are symmetrical and may contain hundreds of thousands of stars; as seen from Earth they are so condensed towards their centres that they are difficult to resolve into individual stars. Even so, the danger of stellar collisions occurring remains slight, but to an inhabitant of a

See also

1


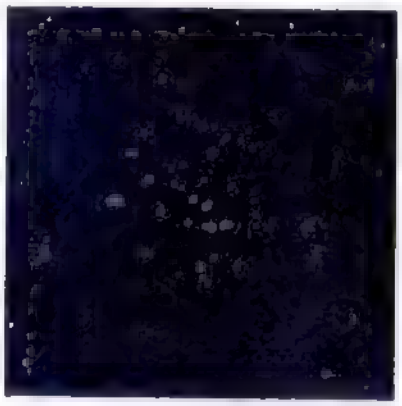
1 The colour-magnitude graph of open clusters shows absolute magnitude (Mv) against colour index (B−V = photographic magnitude minus visual magnitude). The age (years) is given on the right hand scale.

2


2 The twin clusters in Perseus were catalogued by Dreyer as NGC 869 and 884; they are also known as Chi and h Persei. Each cluster is 75 light years across and contains about 350 stars.

3


3 The Hyades round Aldebaran (not itself a cluster member), are easily visible to the naked eye. The group extends in a kind of V formation away from Aldebaran. The individual stars are easy to make out and one of them, Theta Tauri, is a naked eye double. The Hyades are so scattered that the group is not nearly as spectacular as the Pleiades. The best view is obtained with binoculars or with a telescope of low magnification and a wide field.

4


4 Praesepe, in Cancer, is an example of an open cluster visible to the naked eye. It is 525 light years away and has long been known. It is not a condensed cluster and lies well away from the galactic plane.

5


5 The Pleiades cluster (member 45 in the Messier catalogue) is 410 light years away. At least seven of its stars are visible to the naked eye, and the total number of members is about 500. The leading stars are hot and white.

planet that was moving round a star in such a region, the night sky would be scattered with thousands of stars that would be shining more brightly than Sirius or even the full Moon as it appears to us.

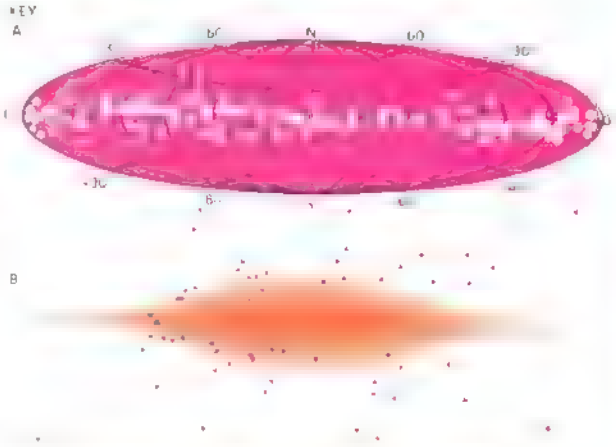
The distribution of the globular clusters is not uniform over the whole sky, they surround the galactic centre, so that from Earth they are only seen in the direction of the centre. The distances of globular clusters have been measured by using the RR Lyrae variables contained in them. Because all RR Lyrae stars have similar periods and luminosity, their distances can be calculated more easily, enabling more effective measurements of the globular clusters themselves. It was by this method that Harlow Shapley, more than 60 years ago, was able to work out the size of our Galaxy from his studies of globular clusters. Because of their remoteness they form an "outer framework" to the main galactic system [Key].

The brightest globular clusters, Omega Centauri and 47 Tucanae, lie so far south in the sky that they cannot be seen from Europe or most of North America. Both are easily visible with the naked eye and Omega Centauri in particular is a superb sight through a telescope, it is resolvable to its centre. In the north, the best example is M13 in Hercules [9], which lies at a distance of 26,700 light years and is about 100 light-years in diameter. It is just visible to the naked eye and binoculars show it well.

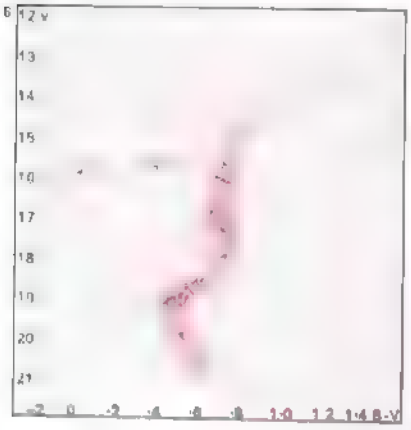
Globular clusters belong to the galactic halo and move round the nucleus in highly inclined, eccentric orbits. They are stable and are not subject to disruption in the same way as open clusters, they are made up of Population II objects, so that their brightest stars are mostly of late spectral type.

Moving clusters

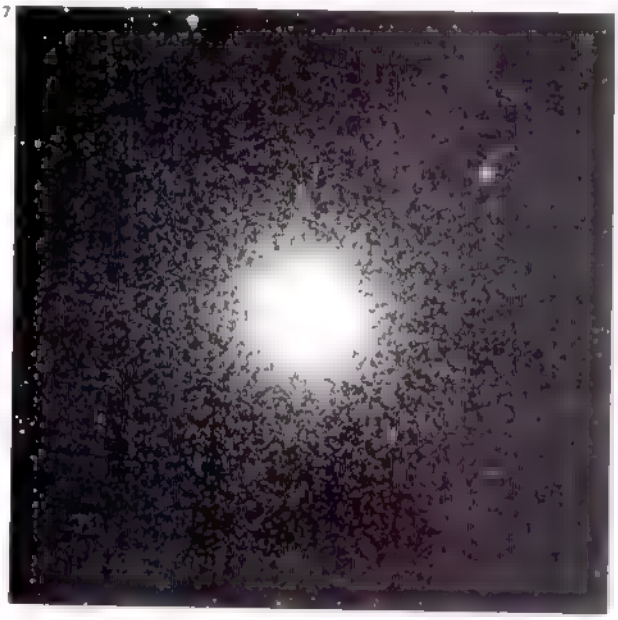
In addition to the open and globular clusters, there are the so called moving clusters, whose members are widely separate but move through space in the same direction and with the same velocity. Hot, luminous O and B-type stars form "stellar associations" of which nearly 100 are now known. One association of this kind is centred on the Orion Nebula.



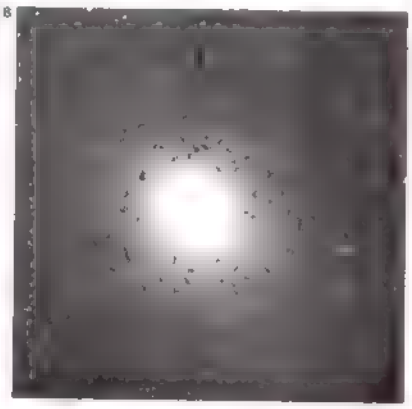
Two main types of stellar clusters are found in and around our Galaxy. Open or loose clusters [A] are composed of Population I stars and they lie near the main plane of the Galaxy, although there are a few exceptions, clusters that are well away from it. An example is the old cluster M67. The open clusters are therefore a part of the general rotation of the Galaxy round its nucleus. Globular clusters [B] which are made up of Population II stars are distributed throughout the galactic halo. The position of the Sun is indicated by the small open circle.



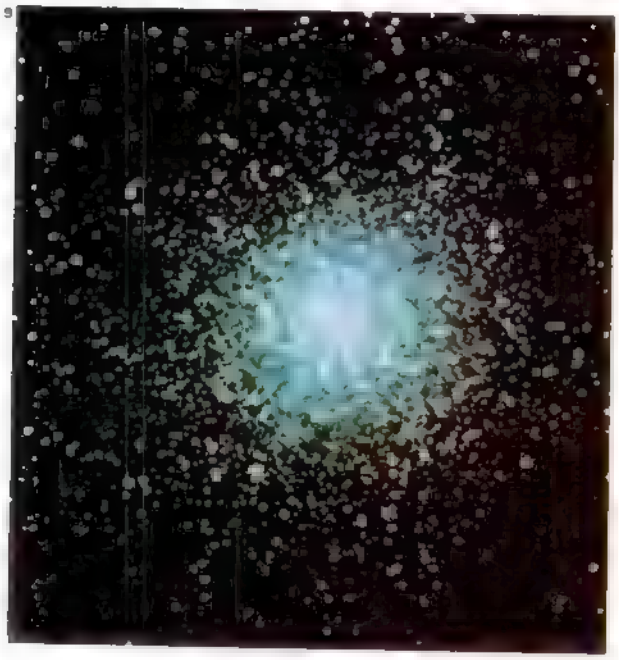
6 The globular cluster M3 in Canes Venatici (the Hunting Dogs), is 48,500 light-years from us. In this colour-magnitude diagram apparent magnitude, v, is plotted against the colour index (B–V). The RR Lyrae variables, used to measure the distance of the cluster, lie at v = 15.7 – the gap in the horizontal branch.



7 The globular cluster M3 in Canes Venatici contains more than 44 000 stars of magnitude 22.5 or brighter within an 8' radius of the centre. It is typical of its kind and quite symmetrical.

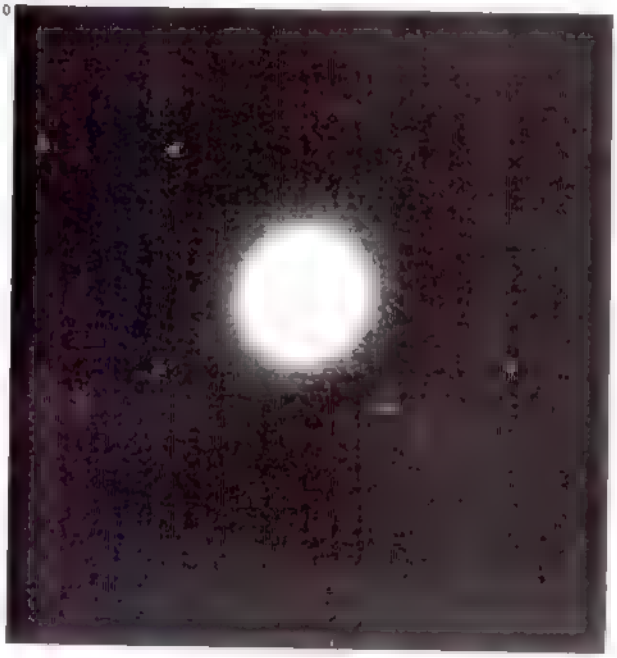


8 The globular cluster M5 in Serpens was discovered in 1702 and is a bright telescopic object. It is exceptionally rich in RR Lyrae variables and more than 100 have been discovered in it.



9 M13 in Hercules is the finest globular cluster in the northern sky and is found between the stars Zeta and Eta Herculis. It is 26 700 light years away from Earth. Surprisingly it is poor in RR Lyrae variables and fewer than 20 have been discovered, as against over 100 in other clusters.

10 Globular clusters attend other galaxies as well as our own, there are many for instance, round the Andromeda Spiral. This photograph shows the giant elliptical galaxy M87 around which some 1 000 clusters have been detected. These appear as small dots on the picture. M87 is at a distance of 40 million light-years from Earth.



Our Galaxy

The Solar System, with the Sun at its centre is an insignificant part of a local system of about 100,000 million suns called the Galaxy. In relation to those other stars, the Sun is neither an exceptionally luminous nor an exceptionally dim star.

The Sun is without doubt older than the Earth and its age cannot be less than 5,000 million years. The Galaxy itself is presumably much older, although it cannot be claimed that we yet have any clear knowledge of its early history. The Galaxy is a flattened system [1] and when we look along its main plane we see many stars in almost the same line of sight. That is why the stars in the Milky Way seem so closely crowded together that they appear almost to touch [Key, 6, 7].

Early observations

The Milky Way is a dominant feature of the night sky, although unfortunately city dwellers never see it clearly because its soft radiance is apt to be swamped by the glare of street lights. One of the earliest and best descriptions of it was written about AD 150 by Ptolemy of Alexandria, the last great classical astronomer and mathematician.

The Milky Way", he wrote, "is not a circle but a zone, which is everywhere as white as milk and this has given it the name it bears. Now this zone is neither equal nor regular anywhere, but varies as much in width as in shade or colour, as well as in the number of stars in its parts and in the diversity of positions, also in some places it is divided into two branches, as is easy to see if we examine it with a little attention."

Most countries have legends about the Milky Way, but its nature was not known until Galileo (1564–1642) first examined it with a telescope in the winter of 1609–10. He found it to be made up of "a mass of innumerable stars", an observation that may be checked even with binoculars.

The first man to record the approximate shape of our Galaxy was William Herschel (1738–1822), who compared it to "a cloven grindstone". Herschel is best remembered for his discovery of the planet Uranus in 1781 although his major contributions were in the field of stellar astronomy. In particular he considered whether the starry or resolvable nebulae, such as that in the constellation of Andromeda, might be separate star systems well beyond our own.

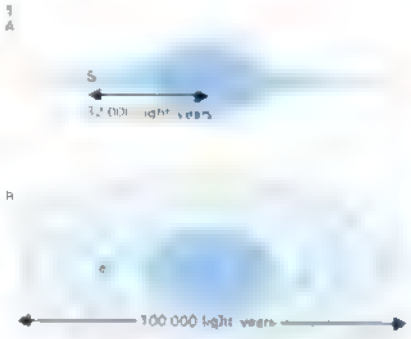
We can never see the centre of the Galaxy because of the obscuring effect of the interstellar materials. Present knowledge is derived mainly from radio astronomy, by which the centre can be located. It lies beyond the glorious star clouds in Sagittarius where the Milky Way is particularly thick with stars [4]. It has been suggested that there may be a quasar there or even a black hole, but these speculations rest on uncertain evidence. Certainly radio waves come from the galactic centre and it was the source of the first radio waves from the sky ever to be detected by Karl Jansky (1905–50) – in the early 1930s.

The form of our Galaxy

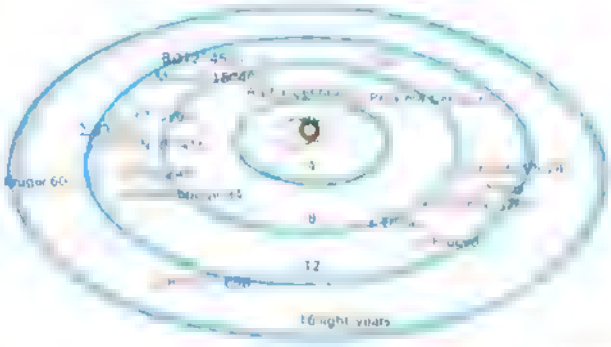
During World War I, Harlow Shapley (1885–1972) in America measured the size of our Galaxy (from his studies of RR Lyrae variable stars in globular clusters). He also proved that the Sun, together with the Earth and the other members of the Solar System

CONNECTIONS

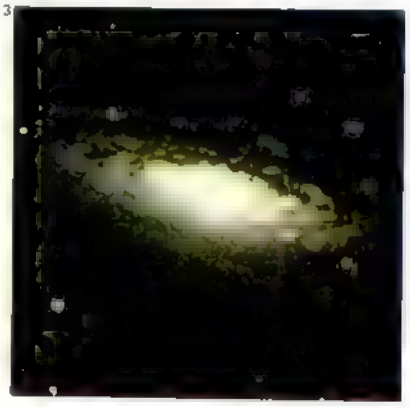
See also



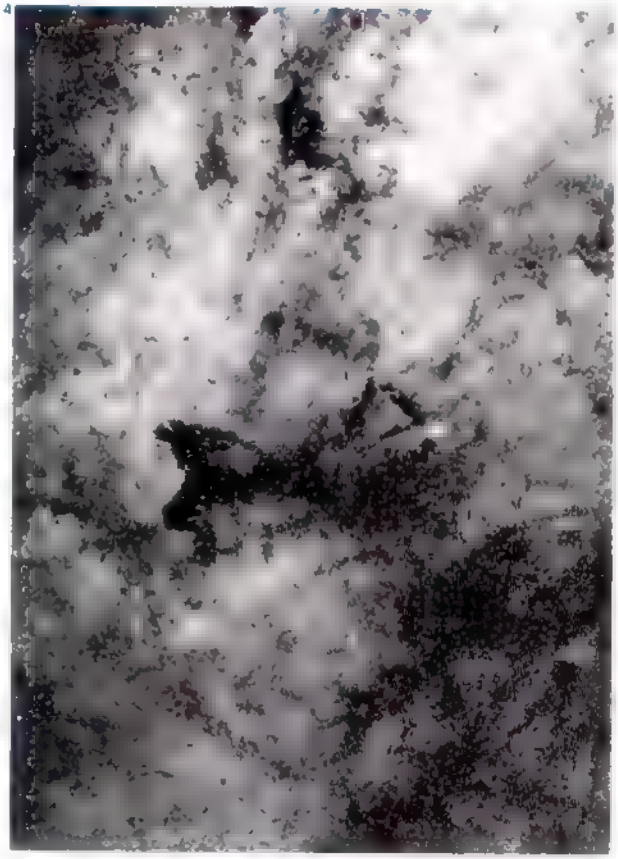
1 The shape of the Galaxy would look different if viewed from separate vantage points well out in space. In an edge-on view [A] the shape is flattened with a pronounced nucleus. (S indicates the position of the Sun.) As seen from an angle [B], the general form of the Galaxy remains clear but the spiral arms are now displayed. The Galaxy is a rather loose spiral, but the arms can be clearly seen.



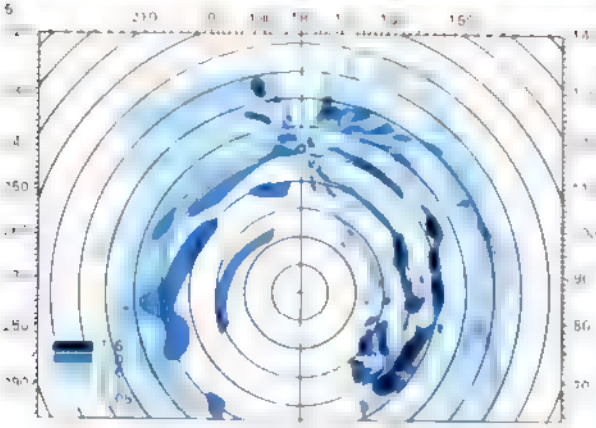
2 Stars within a radius of ten light years from the Sun make up only a small area of the Milky Way. The two ellipses around the Sun indicate the possible maximum distances of orbiting comets. Most of the nearby stars are faint red dwarfs; there are also some white dwarfs (such as the companion of Sirius). The only stars more luminous than the Sun are Sirius, Procyon and Alpha Centauri.



3 Galaxy NGC 7331 in Pegasus is similar in size and mass to our own, although its spiral arms are wound more tightly. It contains some 100 000 million stars. Only when the existence of galaxies outside our own was established could a true idea of the status of our Galaxy be worked out; even in 1920 it was believed that our Galaxy was unique and that the other spiral galaxies were contained inside it.



4 The centre of the Galaxy lies beyond the star clouds in Sagittarius which are seen in this photograph taken with the 48in (122cm) Schmidt telescope at Palomar. In addition to the star clouds, the photograph shows a considerable amount of dark nebulosity: stellar matter which betrays its presence by blotting out the light of stars beyond. This is the richest part of the entire Milky Way area.



5 Radio maps like this one have been drawn to show the distribution of the clouds of cold rarefied hydrogen in the plane of the Galaxy; the clouds are indicated in blue and contours have been drawn which are in accordance with a density scale giving the average number of atoms per cubic centimetre. The map is incomplete but the indications of spiral structure are unmistakable. These regions have a high proportion of Population I stars.

ies well away from the centre, the modern estimate of its distance from the galactic nucleus is 32,000 light-years (rather greater than was believed until recently). However, the comparative size and structure of the Galaxy was uncertain and became clear only with the work of Edwin Hubble (1889–1953), who showed in the 1920s that the spiral galaxies [3] are external systems, of the same basic type as our own.

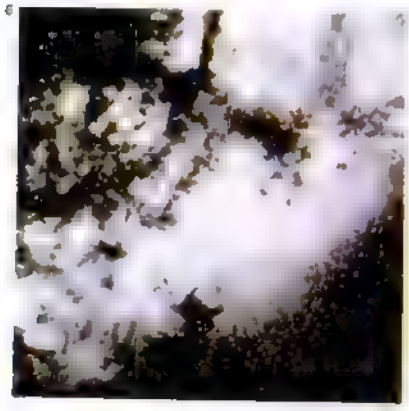
If so, then there seemed no reason to doubt that our own Galaxy must also be spiral (although because the Earth is situated inside it, the spiral effect is naturally lost). The distribution of bright stars ("Gould's Belt") provided some sort of confirmation but the final proof came from radio astronomy. During World War II Hendrik van de Hulst (1918–) and his colleagues in The Netherlands calculated that the clouds of cold hydrogen spread through the Galaxy should radiate at a wavelength of 21cm and in 1951 E. Purcell and H. Ewen, in the United States, showed that this is what does happen. When the positions and the movements of these hydrogen clouds were worked out, an unmistakable spiral structure was found [5]. It has also been established that the Galaxy is rotating, not as a solid body, at one rate but showing differential rotation. In the neighbourhood of the Sun, the revolution period is approximately 225 million years, known unofficially as the cosmic year. One cosmic year ago, the Earth was at the beginning of the Triassic period, when the giant reptiles were replacing amphibians as the dominant life form.

The relative size of the Galaxy

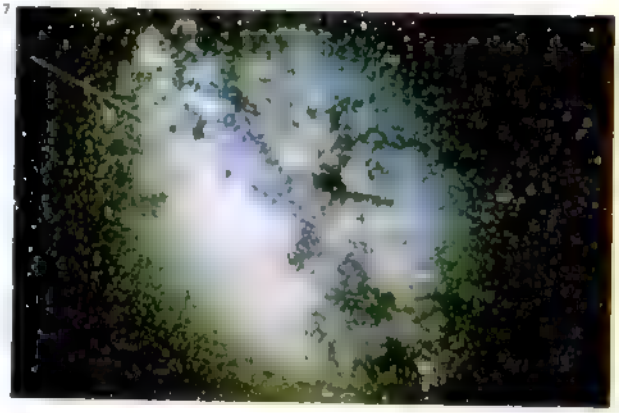
It used to be thought that the Galaxy must be exceptionally large, but this illusion, too, has now been shattered. Its size is above average but some other known systems are decidedly bigger, including the Andromeda Spiral M31, in our local group.

Nowadays the term "Galaxy" is taken to refer to the star system and Ptolemy's "Milky Way" to the luminous aspect in the sky. In appearance, the Milky Way is extremely beautiful, it is particularly rich in areas such as Crux, Cygnus and the Scorpio–Sagittarius region.

NGC 7000 has been given the unofficial title of the "North American Nebula" because of its resemblance to that [illegible] in the constellation of Cygnus, 1,000 light-years away from Earth. This photograph was taken with the 48 in (122cm) Schmidt telescope at Palomar. The nebula is associated with the exceptionally luminous supergiant Deneb or Alpha Cygni. The comparatively very dark areas seen in the photograph are caused by an intervening cloud of opaque dust which cuts out the light of the nebula as well as that of the background stars. The nebula is one of the richest areas of the Milky Way – the galactic plane of our Galaxy.

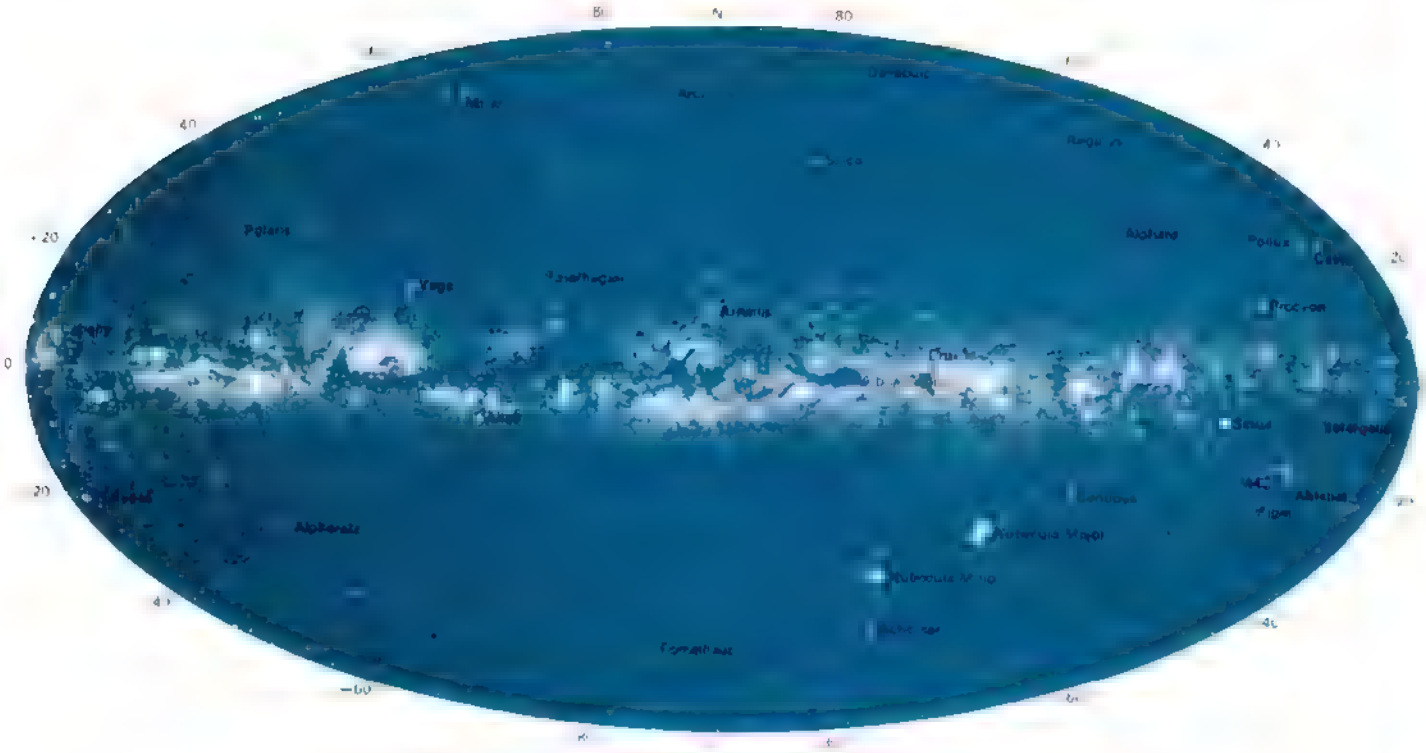


6 Myriads of stars can be seen in this photograph, taken at Palomar although it covers only a small part of the Milky Way. Many of the stars are much more luminous than our Sun.



7 These star fields in the Milky Way were photographed at the Naval Observatory, Flagstaff, Arizona. The trail of the balloon satellite Echo I is seen as it crossed the field of view during the exposure.

8 The Milky Way, as mapped by Martin and Tatiana Keskula at the Lund Observatory in Sweden, has co-ordinates referring to galactic latitude and longitude measured from the galactic plane. The zero point for longitude is the intersection between the galactic plane and the celestial equator near the borders of Aquila and Serpens. The north galactic pole is in Coma Berenices, the south in Sculptor.



The local group of galaxies

Galaxies tend to occur in groups that are usually termed clusters – although there must be no confusion with the open and globular clusters of stars that occur in our own Galaxy and others. Many clusters of galaxies are known, some of which include hundreds of members. It is therefore not surprising to find that our Galaxy is a member of such a system, known generally as the local group.

The local group [Key] also includes the Andromeda Spiral, the Triangulum Spiral, the two Clouds of Magellan and more than two dozen smaller systems that are classed as dwarf galaxies. There may also be two large extra members, known as Maffei 1 and Maffei 2 in honour of their Italian discoverer Paolo Maffei, but neither of these systems can be seen properly because they lie inconveniently close to the plane of the Milky Way and are heavily obscured by dust inside our Galaxy. Their nature is thus uncertain.

A stable system

It is generally agreed that the universe is expanding and that all the galaxies outside the local group are receding at various rates. The members of the local group, however, are not moving away from our own Galaxy (indeed, the Andromeda Spiral actually shows a movement of approach, although for the most part this is because of the Sun's own motion around the galactic centre). The local group makes up a stable system.

In attempting to determine the membership of the group, theorists face a serious initial difficulty. The distance of the Andromeda Spiral [1, 2] was originally estimated as being 750,000 light-years. But this estimate produced a number of anomalies: the globular clusters surrounding the spiral were calculated, at that distance, to be systematically different in size from similar globular structures in our Galaxy. Equally unexpectedly, RR Lyrae stars could not be found inside the Andromeda Spiral. Although RR Lyrae stars are less luminous than Cepheid variables, they should still have shown up quite well had the spiral been only 750,000 light-years away. Even a subsequent increase in the estimated distance of the Andromeda Spiral to 900,000 light-years did not resolve these anomalies.

The distances of the galaxies had been calculated by reference to the period and luminosity of the Cepheid variables. In 1952 Walter Baade (1893–1960) announced that the accepted Cepheid scale was inaccurate. There are, in fact, two kinds of Cepheids, one much more powerful than the other. The Cepheids in the Andromeda Spiral had been taken as being less luminous but they were really members of the highly luminous class, which meant that they were much more remote than had been believed. Today, the Spiral is estimated to lie at a distance of 2.2 million light-years.

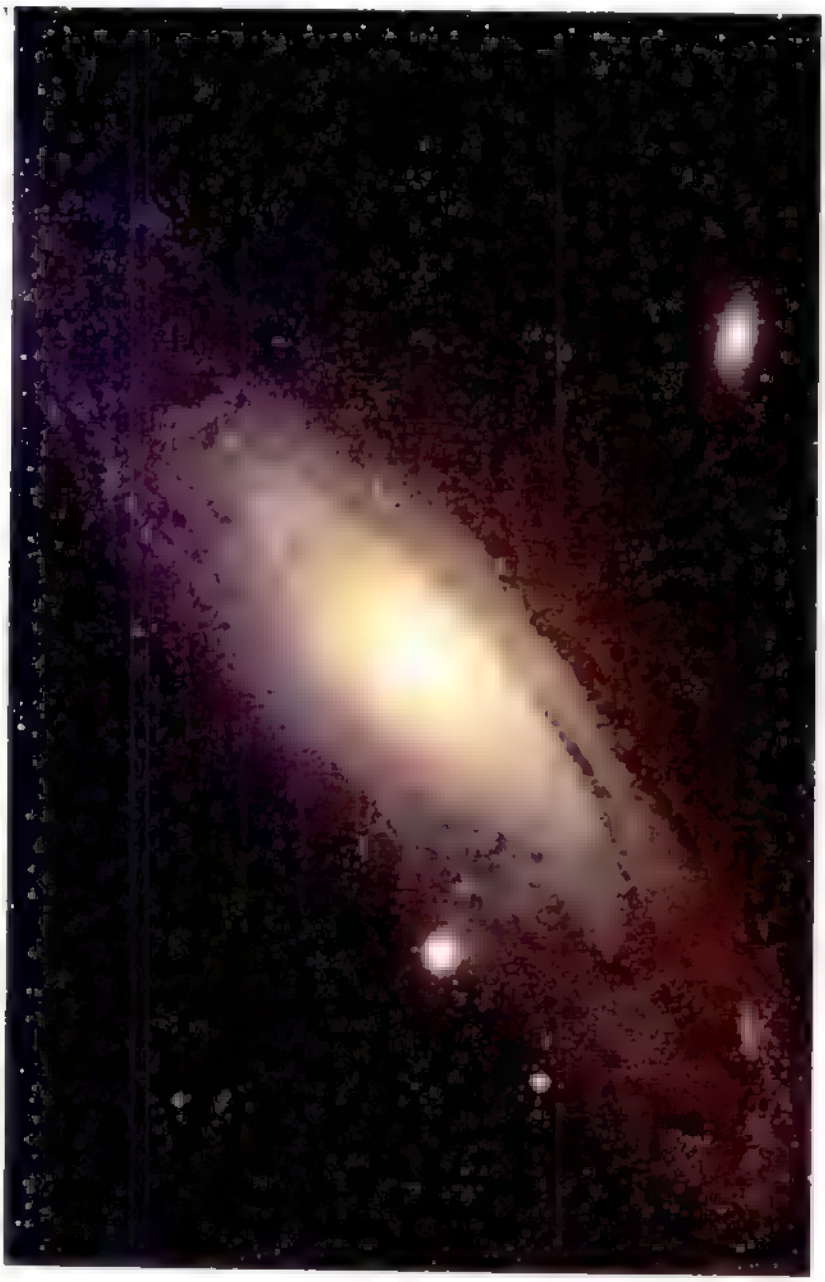
Distance and relative size

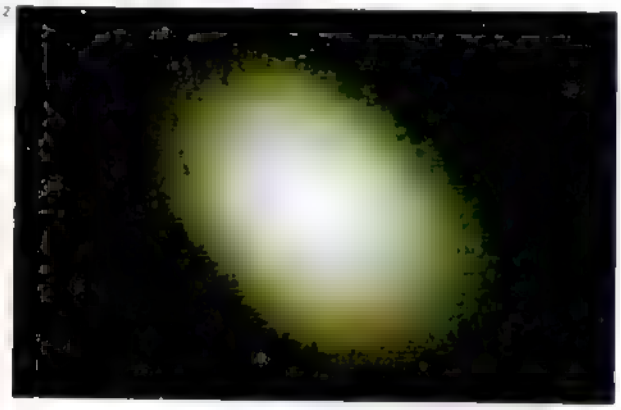
The Andromeda Spiral is the largest member of the local group, with our own Galaxy a rather poor second. Next comes the Triangulum Spiral [4] and then the two Clouds of Magellan or Nubeculae, which are too far south to be visible from Europe or the United States. The remaining galaxies in the local group are much smaller and less rich.

The Clouds of Magellan [5, 6] look rather like detached portions of the Milky Way, but

CONNECTIONS

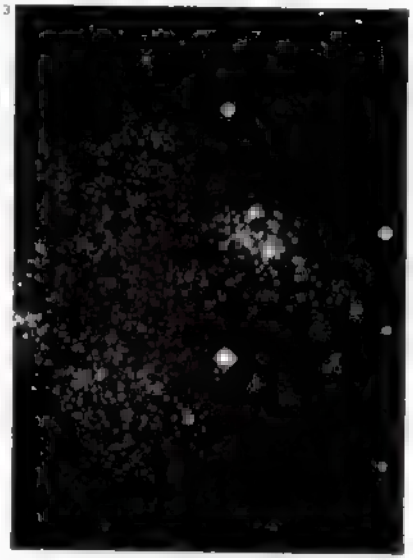
also





1 The great spiral in Andromeda has been known for many centuries and was noted by the Arab astronomer Al Sufi in the tenth century. It was first described telescopically by Simon Marius (1570–1624), a contemporary of Galileo, as looking 'like the flame of a candle seen through horn'. It is distinctly visible to the naked eye under good conditions, but even through a telescope of considerable size it is disappointing, appearing to be nothing more than an elongated blur of light. Part of the reason for this is that the system lies at a difficult angle from Earth. Were it face-on, as is the Whirlpool Galaxy (M51) for instance, it would be more imposing. As well as containing clusters, gaseous nebulae and variables of all kinds, novae have also been seen frequently. In 1885, one supernova, S Andromedae, could just be seen with the naked eye.

2 The nucleus of the Andromeda Spiral, M31, is usually overexposed in photographs to bring out the structure of the spiral arms. Where the exposure is correct (as in this picture) the arms do not show up. The scattered stars are members of our own Galaxy and simply happen to lie in the foreground. The nucleus of M31 contains mostly Population II stars, as do many galactic nuclei.



3 The dwarf galaxy in Sextans (centre) in the local group has relatively few stars. The bright star that appears below it is a foreground star in our own Galaxy.

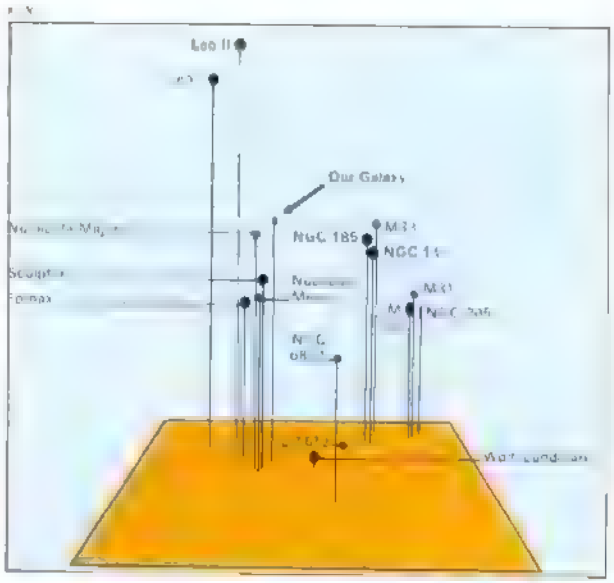
each lies at a distance of about 150,000 light-years from Earth. They are irregular in form (suggestions of spirality in the Large Cloud do not seem convincing) and have been regarded as satellite galaxies of our own Galaxy, though whether they are moving around our Galaxy is extremely problematical. The richer Large Cloud is about 40,000 light-years in diameter and the Small Cloud 20,000 light-years, so that both are much smaller than our Galaxy. The distance between their centres is 75,000 light-years and their genuine association is shown by the fact that both seem to be contained in a common envelope of rarefied hydrogen. Population I characteristics are evident: novae and large gaseous nebulae have been seen in them. Both contain Cepheid variables.

The Andromeda Spiral also has two companion galaxies, M32 (NGC 221) and NGC 205. Both are dwarf elliptical systems, made up of Population II stars. They are easily visible in small telescopes, but do not seem to be of the same class as the Clouds of Magellan. For a long time it was thought that the Clouds of Magellan, the spirals in Andromeda and Triangulum and the companions to the Andromeda Spiral were the only members of the local group, but many smaller systems have now been discovered.

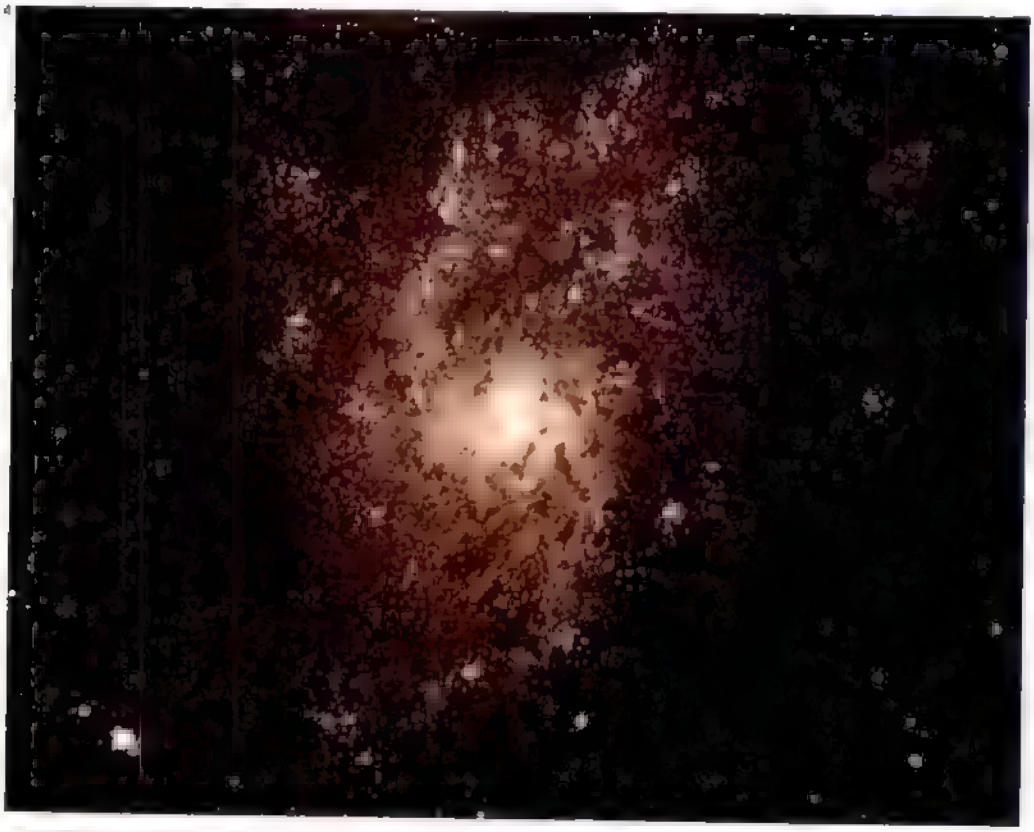
Sparsely populated systems

All these smaller systems are of relatively low mass; for instance, the two dwarf systems in Sculptor and Fornax make up a total of only about one per cent of the mass of our Galaxy. These and similar systems are so sparsely populated that they are not immediately recognizable as galaxies in their own right. They are made up of Population II stars, so the leading stars are old red giants. There is little or no interstellar material prominent enough to be detected from Earth, which suggests that star formation has ceased.

There seems to be no real difference between the local group and many other groups of galaxies known to us. However, the local group is much easier to study because it is in our own part of the universe. The really small galaxies, such as those in Sculptor and Fornax, would not be detectable at all if they lay at distances of millions of light-years.

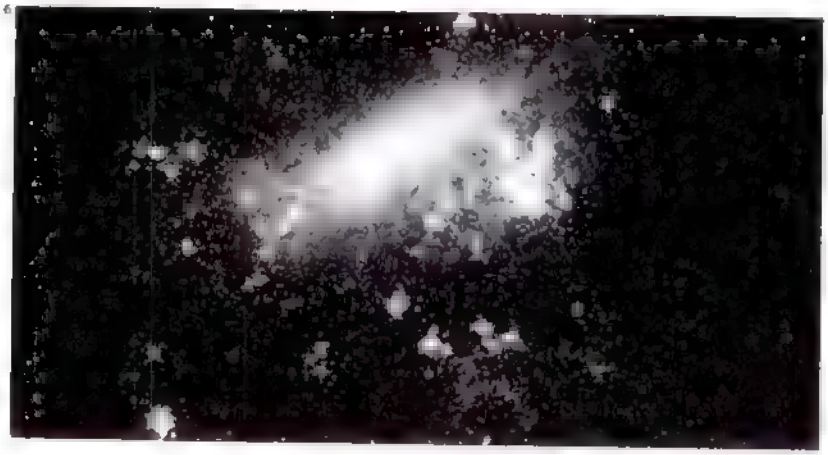


The local group of galaxies is small with fewer than 30 members. Only the Andromeda Spiral (M31) and the Clouds of Magellan, Nubecula Major and Minor, are naked-eye objects

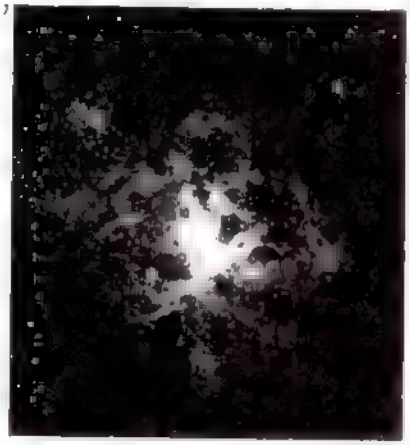


4 The Triangulum Spiral, M33, is the third largest member of the local group and the loose spiral form is well shown. At its distance of 2,350,000 light-years, it is slightly more remote than M31.

5 The Small Cloud of Magellan lies in the far southern constellation of Tucana. It is an easy naked-eye object in a dark, clear sky. It contains many variable short-period stars.



6 The Large Cloud of Magellan is irregular in shape and contains many stars. Part of it lies in the constellation of Dorado, part in Mensa. It is a bright naked-eye object even in moonlight.



7 In the Large Cloud lies the Great Looped Nebula surrounding the star 30 Doradus. This star itself is variable and is the most luminous star known, about a million times as powerful as the Sun. But it cannot be seen without optical aid.

Types of galaxies

As can be seen from even a cursory examination of photographs, galaxies assume various forms [Key]. There are galaxies that show a spiral pattern, some of them loose, some tightly coiled. Among these are the barred spirals, in which the arms seem to protrude from the ends of a "bar" apparently through the centre of the system. Other galaxies appear elliptical, ranging from extremely long, narrow systems to shapes that are almost circular as seen from Earth. Finally there are irregular galaxies, with no definite shape at all. Most of the dwarfs belong to this category, but there are larger irregular systems such as M82, a radio source lying within Ursa Major.

The Hubble classification

The study of galaxies entered its modern phase in the early 1920s, when work carried out by Edwin Hubble (1889–1953), using the 100in (254cm) Hooker reflector telescope at Mount Wilson in California, definitely confirmed the existence of external systems that were not outlying parts of our own Galaxy. Hubble established a method of classification that has served as the basis of later, more complicated classifications [1]. Hubble's system distinguished three basic types of galaxies, the spirals, the ellipticals and the barred spirals. The irregular galaxies were not classified separately although they were recognized. It was clearly tempting to regard the Hubble classification as an evolutionary sequence. However it is purely a classification intended to recognize increasing degrees of flattening, the elliptical systems are in fact spheroidal – they look elliptical only in projection. Present knowledge is so limited that most astronomers regard any overall evolutionary sequence with a certain degree of scepticism.

There is still little idea of how spiral arms are formed, although it has definitely been established that the stars in the disc of a spiral galaxy mostly revolve about the centre in approximately circular paths in the same direction. The spiral pattern appears to revolve also in the same direction with the arms "trailing". There is uncertainty about whether or not the spiral arms – apparently some sort of wave moving through the stars and gas – are long-lasting on the cosmic scale. Spirals, with their numerous hot Population I stars and their interstellar matter, appear to be less advanced in their evolution than the elliptical galaxies, where the leading stars are red giants of late spectral type and where there is comparatively little nebular material to be observed.

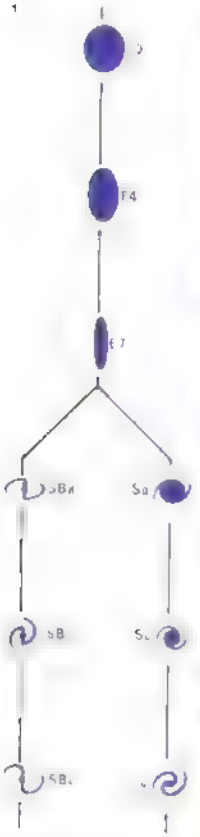
In what are known as the Seyfert galaxies – first noted by Carl Seyfert (1911–60) in 1943 – the nuclei are almost stellar in appearance and the spiral arms are relatively obscure and tightly wound. Seyfert galaxies emit radio waves and great disturbances may well be in progress there. A particularly good example of a Seyfert galaxy is M77 in Cetus, which has a total mass estimated at about 800,000 million times that of the Sun.

The Hubble constant

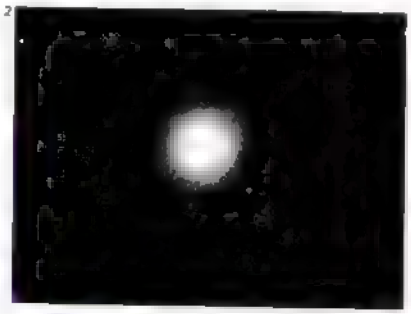
Even before Hubble's researches had proved that the galaxies lie well beyond our own system, it had been found that the 40-odd galaxies for which suitable spectra had been obtained were apparently receding. This was established by the Doppler shift [11]. If a

CONNECTIONS

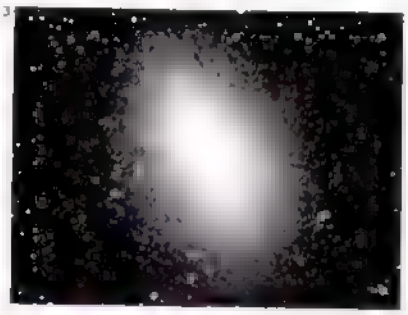
See also



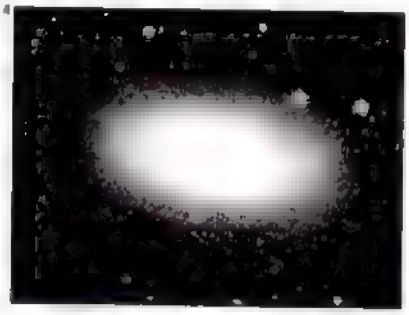
1 The Hubble system classifies three types of galaxies. Ellipticals range from E0 to E7, those of E0 appear spherical and, as a result, look misleadingly like globular clusters, those of E7 are decidedly elliptical. Spirals are of types Sa (large nucleus and tightly wound arms) Sb (smaller nucleus, looser arms) and Sc (small nucleus, loose arms). barred spirals are divided into SBa SBb SBc. Hubble's system is still used



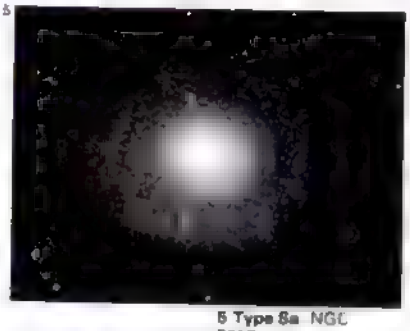
2 Type E0 M87 in Virgo is a symmetrical system and a powerful radio source



3 Type E4 dwarf galaxy NGC 147 in Cassiopeia contains Population II stars

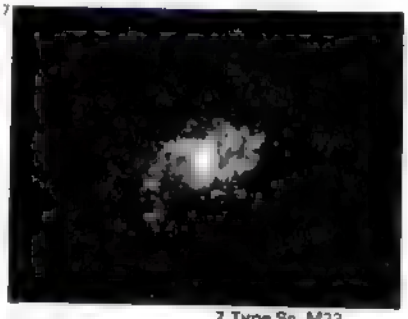


4 Type E6 NGC 205 is a smaller companion of the Andromeda Spiral

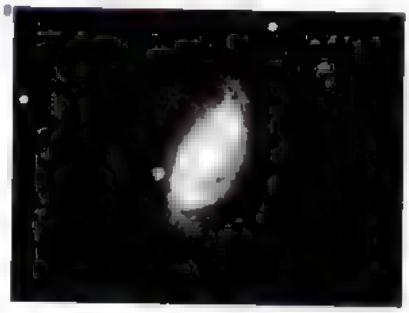


5 Type Sa NGC 7217, a spiral in Pegasus, has arms that are symmetrical and tightly wound.

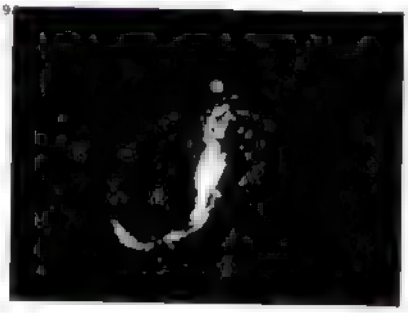
6 Type Sb M81 (NGC 3031) in Ursa Major is seen at a narrower angle than NGC 7217



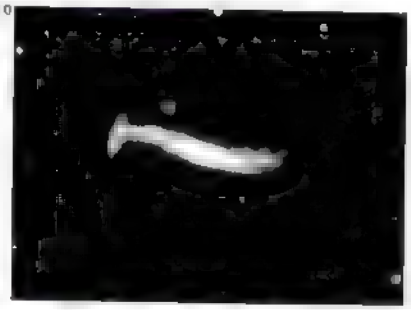
7 Type Sc M33 (NGC 598), the Triangulum Spiral, has a less defined nucleus and less clear arms



8 Type SBa NGC 3504 in Leo Minor has spiral arms extending from the ends of a bar



9 Type SBb NGC 7479 in Pegasus has a pronounced bar with arms clearly extending from its ends



10 Type SBc in the Hercules cluster has a dominant bar and the arms are little more than extensions

galaxy is receding, the lines in its spectrum are shifted towards the red or long-wave end. The farther away a galaxy lies, the greater the speed at which it is receding.

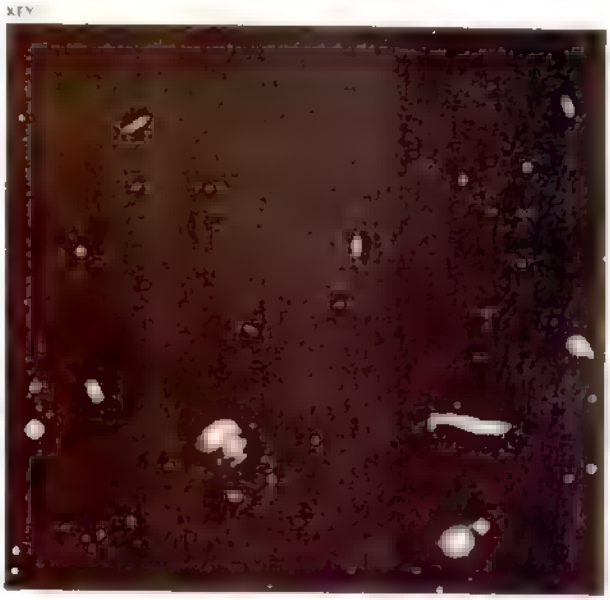
Hubble himself established that there is a definite empirical relationship between distance and the speed of recession. In essence the speed of recession is proportional to the distance, the factor of proportionality is known as Hubble's constant.

Measuring distances

Measurement of the distances of galaxies cannot be precise. For nearby systems – those of the local group, and even beyond – the period-luminosity relation for Cepheid variables can be used. This method for establishing the distance of Cepheids, which can then be used as "standard candles" for other stellar distance measurements, seems to be reliable now that the difference between types of variable stars has been clarified. Cepheids are powerful stars that can be observed at distances of up to several millions of light-years.

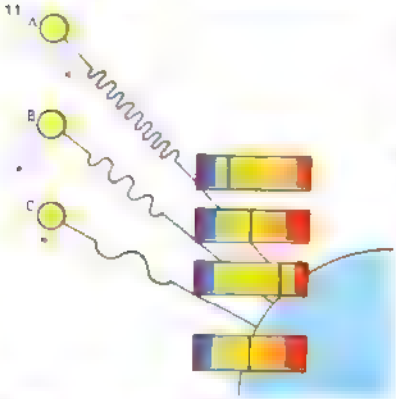
Supergiant stars, however, are more powerful than the Cepheids and it seems probable that the brightest supergiants in our Galaxy are more or less equal to the supergiants in other galaxies, so that they too can be used as indicators – although the results are probably less accurate than with the Cepheids. This method works for distances up to about 40 million light-years. There, in the Virgo cluster of galaxies, are galaxies of all kinds, including spirals. Having calculated their sizes it is possible to use them in the same way as the Cepheids and the supergiants, although accuracy is again less. Supernovae can also be used when they appear in remote systems. For the extremely distant galaxies there is as yet no means of measuring a distance independently of the red shift, but if the red shift can be measured the distance may be inferred (assuming that Hubble's law is valid).

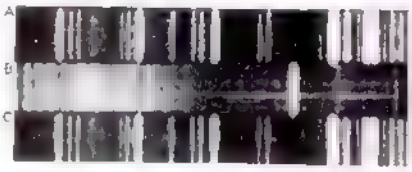
Galaxies that lie beyond the local group yield little detail even when viewed with large telescopes. Only detailed photography can reveal the diverse and fascinating nature of more distant star systems scattered throughout the universe.



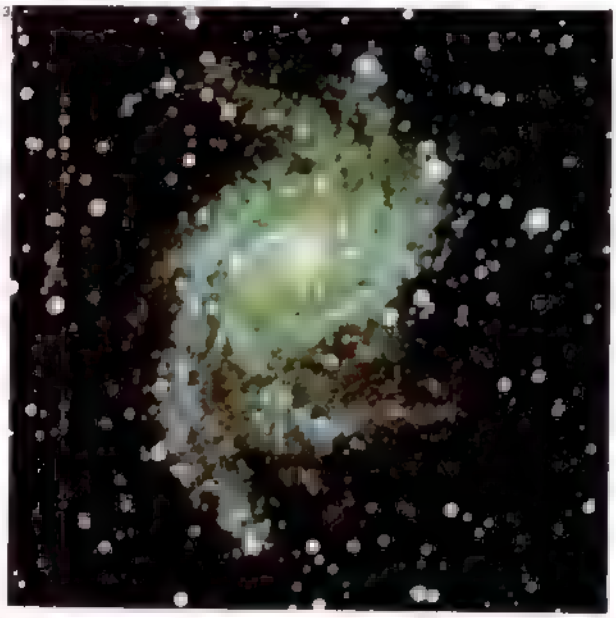
These galaxies in Hercules include spirals, barred spirals and elliptical systems. They form a genuine cluster.

11 The Doppler shift occurs when, from an approaching light source (A), more waves per second enter the eye than would enter from a source without relative motion (B), and the wavelength shifts towards the violet end of the spectrum. Without relative motion (B), the light is unaffected. Recessional velocity (C) produces an apparent increase in wavelength and a spectral red shift.

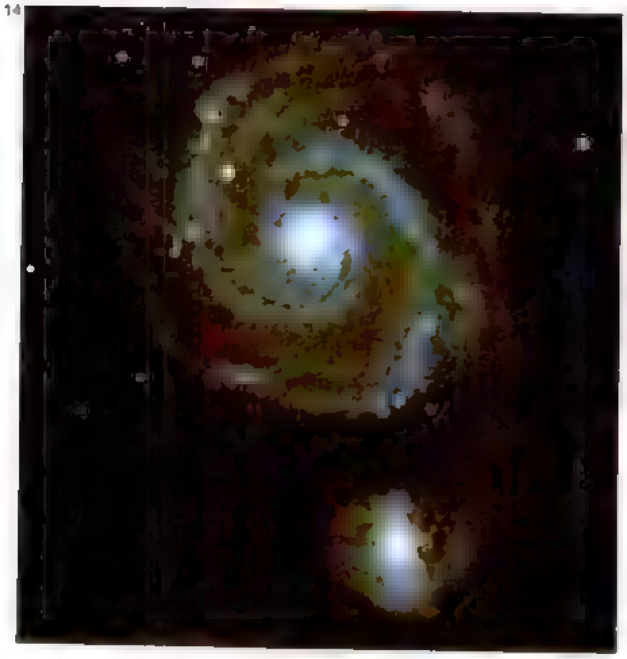


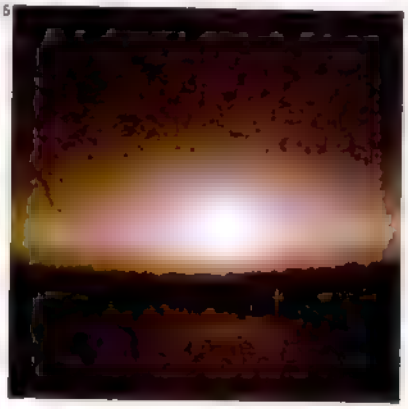


12 One of the most remote galaxies is 3C 295 (arrowed). Its spectrum (B) is compared with laboratory spectra (A) and (C). The position of the white line, shifted to the right, shows that 3C 295 is red-shifted.



13 NGC 5946 is a loose spiral galaxy. The nucleus is well marked but not large compared with the spiral arms, which are much less pronounced and tightly wound than those of galaxies classed Sa.





14 The Whirlpool galaxy, M51 in Canes Venatici, is 37 million light-years away. It is face-on to us and therefore is excellently displayed. This was the first spiral galaxy to be recognized (by the Earl of Rosse in 1845).

15 The Sombrero Hat galaxy, M104, part of the Virgo cluster, is at a distance of 41 million light-years.

16 Spiral galaxy NGC 253, in Sculptor, is almost edge-on to Earth. As a result, the spiral form is not well displayed.

Radio galaxies and quasars

Some galaxies are powerful sources of radio waves as well as light. These are known as "radio galaxies". No doubt all galaxies send out long-wavelength radiations because of the supernova remnants and other discrete radio sources inside them, but the energy of radio galaxies is of an entirely different order of magnitude.

Problems of radio galaxies

A typical galaxy with a strong radio source is M87 in the famous Virgo cluster, approximately 60 million light-years from Earth. A curious jet issues from it which looks as though it may be composed of material being ejected at high velocity. M87 sends out about 10,000 times as much energy at radio wavelengths as it might be expected to do. Some other radio galaxies are much more remote: for instance Cygnus A, the first radio galaxy identified optically (in 1954), is 700 million light-years distant (assuming a Hubble constant of 50km/sec/megaparsec (a parsec is the distance at which a star would show a parallax of one second of arc; it is equal to 3.26 light-years).

Many theories have been proposed to explain the radio emissions from these exceptional galaxies. Originally it was thought that there might be not one galaxy, but two – in fact, two separate systems passing through each other in opposite directions. If so the individual stars would seldom collide, but the interstellar matter would be in collision throughout the encounter and would – it was suggested – produce the radio emission that is recorded. Certainly radio galaxies such as Centaurus A give the impression of being compound. Cygnus A, like many other radio galaxies, manifests two powerful centres of radio emission straddling the optical image with a weak radar source coinciding with the optical object.

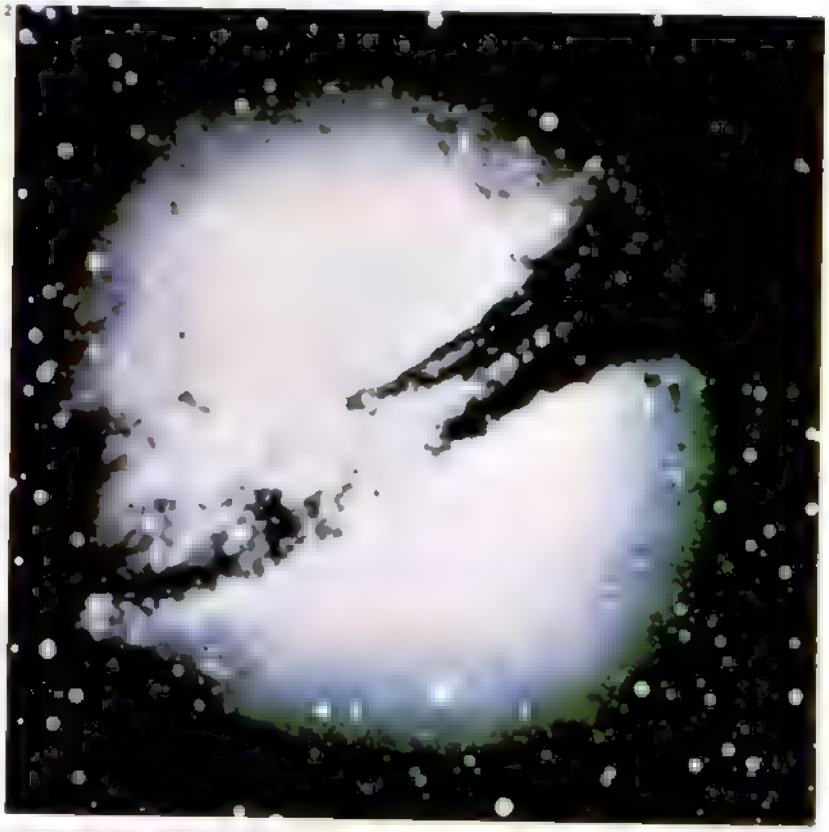
Further research showed that collisions could not produce nearly enough energy to explain the observations and the theory of colliding galaxies was abandoned. It now seems that the radio emissions are the result of tremendous explosions inside the galaxies themselves. One excellent example of this is M82, an irregular galaxy in Ursa Major [3]. It has been found that there are huge, intricate gas structures inside the galaxy, moving about at speeds of up to 160km (100 miles) per second, and from the present movements it seems that an outburst took place near the centre of M82 about one-and-a-half million years ago (although since the distance of the galaxy is 10.5 million light-years, the explosion occurred 12 million years ago in our time frame). The radio emission from objects like M82 is believed to be generated by the synchrotron process – that is, energy radiated by the acceleration of high energy electrons in a strong magnetic field. Unfortunately the cause of these outbursts in radio galaxies is not yet understood.

A new object in the sky

In 1960 the search for distant objects in the universe led to a surprising sequence of events. A few objects, which from their radio properties were believed to be distant, were identified with blue star-like objects on photographs obtained with the Palomar telescope. Until 1963 they were believed to be a hitherto unidentified type of star in the Milky Way. In March of that year M. Schmidt,

CONNECTIONS

See also



1 Galaxies NGC 4038 and 4039 in Corvus are classified Sc because each galaxy is a loose spiral. Each is also a radio source. There is no doubt that the two systems are genuinely associated and that they lie at the same distance from the Earth. They give the appearance of being interlocked and it was this particular aspect that led to the theory that radio emissions from galaxies are due to collisions. If this theory had been valid, NGC 4038 and 4039 would have been spectacular examples. However it is now known that the theory of colliding galaxies is wrong. The star to the lower right is in our own Galaxy and hence shows up in the foreground.

2 The radio galaxy Centaurus A (NGC 5128) is now thought to be a single system although it was once believed to be a collision of two galaxies. At its distance of 12 million light-years, it is one of the closest of the radio galaxies and seems to contain an unusual amount of diffused dusty material. The radio sources do not coincide with the centre of the optical object, but lie on either side

3 The irregular galaxy designated M82 is a radio source 10,500,000 light years away. There seem to be intricate hydrogen gas structures of tremendous size moving at speeds of up to 160km (100 miles) per second, and all the indications are that a tremendous explosion took place inside the nucleus of the system 1.5 million years before our present view of it. It is therefore the best known example of an exploding galaxy.

working with the 200in telescope, succeeded in identifying the spectrum of the radio object known as 3C 273 (that is, the 273rd object in the third Cambridge catalogue of radio sources) [4] and at the same time Greenstein and Matthews published their red shift measurements on another blue object – 3C 48. The red shifts were extremely large and as more objects of this type were identified it became clear that a new class of object had been discovered, more remote than any hitherto recognized and in many cases receding with velocity greater than half the speed of light.

Problems associated with quasars

These distant objects, known as quasars, a contraction of their original name, quasi-stellar objects, have provided astronomers with a series of baffling problems. Assuming that the estimates of their distances are correct, a powerful quasar may far outshine a whole galaxy such as our own – and it is difficult to know how a relatively small object can emit so much energy. Radio measurements of the angular diameters of quasars and the rapidity of the light variations found in some of these objects imply that the main output of energy may be located in a region of space that is only a few light-years in width.

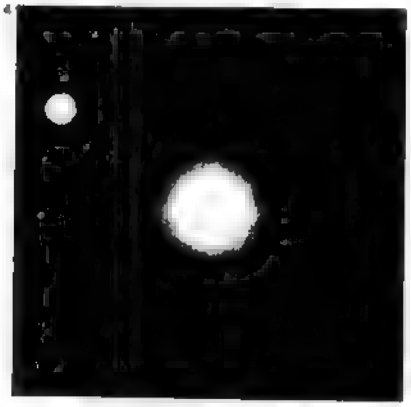
None of the processes of energy production encountered in normal stars or galaxies seems adequate to account for these phenomena and many theories about quasars have been put forward over the years. It was suggested that a quasar might be produced by many supernovas exploding in quick succession, but there seems no reason why such an event should happen, and theories involving the possibilities of anti-matter or black holes are entirely speculative.

It is possible that quasars and certain types of radio galaxies (notably the Seyfert systems – those having bright central nuclei with emission line spectra) are different stages of evolution of the same class of object, but there is still no reliable information. Recently two American astronomers, J. Oke and J. Gunn, have found that the peculiar object known as BL Lacertae [8] may turn out to be a quasar embedded in a normal galaxy.

KEY

Optical telescopes reveal objects in the heavens that emit light, whereas radio telescopes detect their longer wavelength radiation. Each technique can highlight a different feature of the same object. Here the radio map of the Andromeda galaxy (red lines) is superimposed on its optical image. The galaxy is a weak radio emitter, but the map reveals radio sources that are invisible because they emit no light, such as the one near the lower right corner. The wavelengths detected by radio telescopes range from an upper limit of about 30m down to 1cm, below which any incoming radio waves from space are blocked by the atmosphere of the Earth.

4



4 The quasar 3C 273, photographed with the 200in (508cm) Palomar reflector, lies in Virgo and with a magnitude of 13 is visually the brightest of the quasars. It was one of the earliest to be identified, in 1963.

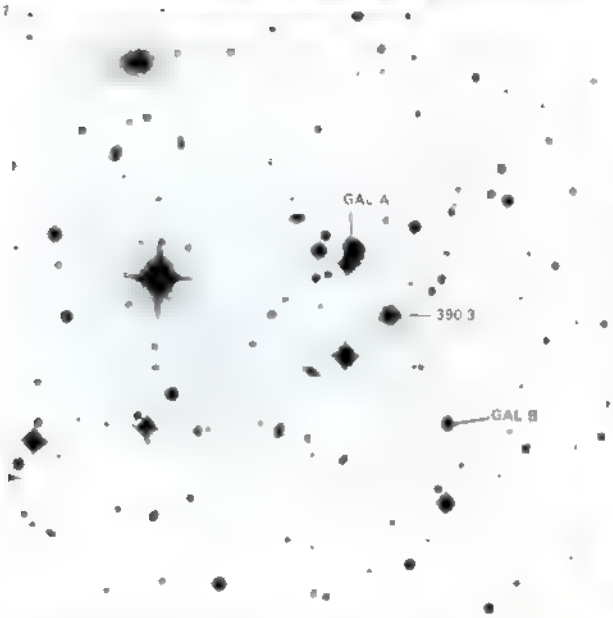
5

5 Quasar 3C 147 has also been photographed from Palomar with the 200in reflector. The quasar (arrowed) looks remarkably like the object just below it, which is an ordinary star in our Galaxy.

6 A

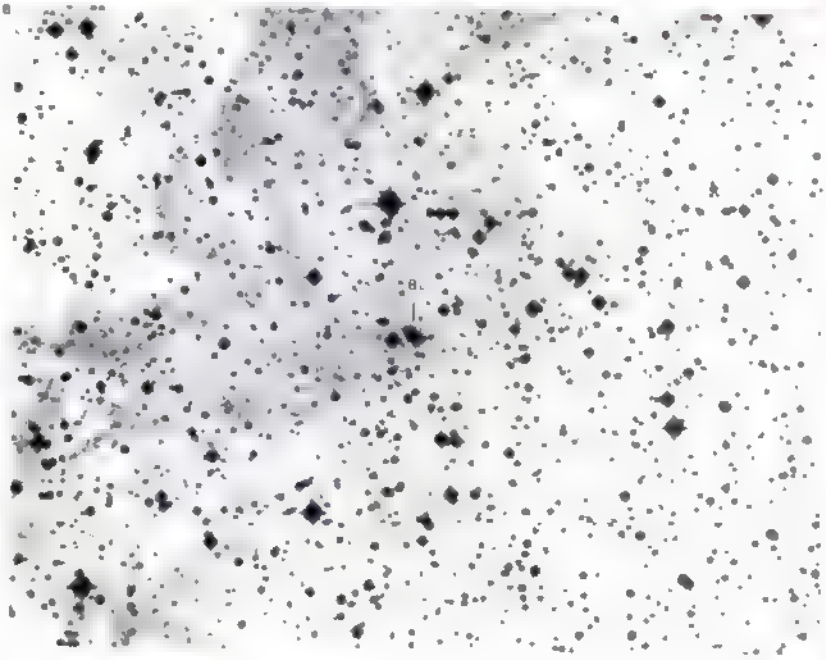
6 These two photographs of the quasar 3C 345 were taken at the Royal Greenwich Observatory in August 1966 (A) and in September 1971 (B). The quasar is arrowed and the decrease in brightness compared with the other stars can be seen quite easily. Quasars show fluctuations over short periods and this is evidence that they must be extremely small compared with galaxies.

7



7 The galaxy marked Gal A is the closest of the three radio galaxies in this photograph and may be associated with the elliptical galaxy to the upper left. 3C 390 3 and Gal B are also radio sources identified optically in this photograph.

8



8 The extraordinary object BL Lacertae, which superficially looks like a star, is variable and near minimum a slight fuzziness is recorded around it. The spectrum is quite featureless, with no bright or dark lines. It has been found that the outer "fuzz" shows a spectrum resembling that of an elliptical galaxy. BL Lacertae is not a star, it is remote and luminous, possibly intermediate in luminosity between a galaxy and a quasar. There is a strong infra-red emission.

The expanding universe

Of all the unanswered questions facing mankind, that of the origin of the universe is one of the most fascinating – and certainly one of the most enigmatic

The Doppler significance

An analysis of the structure of a luminous body will show whether it is approaching or receding. If it is approaching, the wavelength of the light appears slightly shortened and the body appears "too blue". If it is receding, the apparent wavelength is lengthened and the body appears "too red". This is known as the Doppler effect after Christian Doppler (1803–53), the Austrian physicist who first drew attention to it in 1842. The Doppler effect shows up in the spectrum of any body that is self-luminous. If the spectral lines are shifted toward the red or long-wavelength end, then this indicates that the body is receding

The spectrum of an external galaxy is made up of the combined spectra of millions of stars, but the main lines are identifiable and it has been found that apart from the members of our local group of galaxies [1], all the shifts are to the red If these shifts are Doppler effects, it follows that the whole universe is in a state of expansion It has also been found that the farther away a galaxy lies, the greater its red shift – and hence the greater its recessional velocity. This was shown by the work of Edwin Hubble (1889–1953) at Mount Wilson during the years following 1923 when he first showed conclusively that certain nebulae were in fact external galactic systems rather than objects in our Milky Way system

Theories of the universe

Several years before Hubble discovered the observational evidence for the expansion of the universe, a Dutch astronomer, Willem de Sitter (1872–1934), found a solution to the cosmological theory published by the mathematical physicist Albert Einstein (1879–1955) in 1917. Soon afterwards the Russian scientist A. Friedmann (1888–1925) discovered a whole range of solutions to the Einstein equations in which the radius and mean density of the universe varied with time But many problems remain

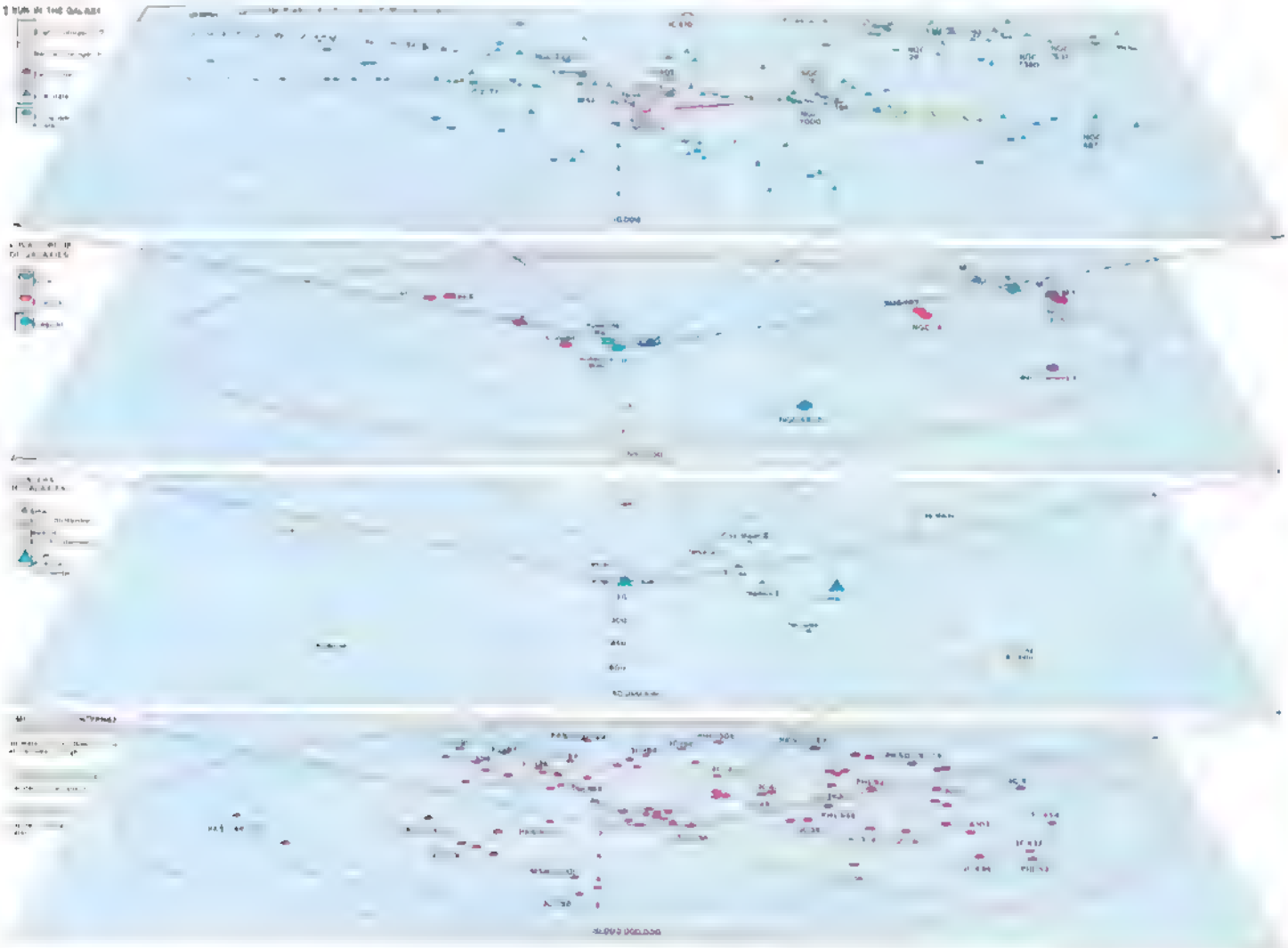
By adjusting the parameters in the equations, the theoretical models predicted either a universe expanding indefinitely as time advanced or one that would eventually collapse. Many distinguished theorists such as Arthur Eddington (1882–1944) and George Lemaître (1894–1966) developed variations of the models of the expanding universe all of which had in common an initial "beginning" in time when the primeval material was compressed into an infinitely small space

In 1946 George Gamow (1904–68) developed the idea (popularized as the "big bang" theory) that this initial state was at an extremely high temperature, resulting in a primordial explosion He also proposed that the common elements were formed from the primeval hydrogen in the first minutes after the beginning of the expansion

The difficult conceptions involved in a beginning of time (and a comparison of the predicted age of the universe and that of the Earth) led Fred Hoyle (1915–) and T. Gold to propose in 1948 that the universe never had a beginning, but was in a steady state of continuous creation. That is, hyd-

CONNECTIONS

See also



1 The immense scale of the universe is shown in these diagrams The region of our Galaxy which can be examined optically is shown first On this scale the entire Solar System would be a microscopic dot In addition to the visible stars there are clusters, stellar associations and gaseous nebulae such as the Rosette Nebula NGC 2237 and the Orion Nebula M42 Distances are given in thousands of light-years, so that the outermost white line represents a distance of 10,000 light-years from the Sun. The local group of galaxies contains more than 24 members, of which the largest are the spiral galaxy M31 in Andromeda, our own, the Triangulum Spiral M33 and the Clouds of Magellan, the companions of our Galaxy, the other members such as Leo I and Leo II are dwarf galaxies. Distances are given in millions of light-years. Two recently discovered galaxies, Maffei 1 and 2, may be members of the local group but are so heavily obscured by dust in the plane of our Galaxy that they are hard to study The galaxies in the local group are not receding from us. The area out to 750 million light-years contains many clusters of galaxies – such as the rich Virgo cluster The region out to 10,000 million light-years cannot yet be studied by either optical or radio methods Even the most remote objects, the quasars, are not as far away as this.

rogen atoms were being continually created and forming into stars and galaxies at a sufficient rate to replace the galaxies that were moving out of the field of view because of the expansion of the universe. During the postwar epoch of radio investigations of the distant parts of the universe great arguments arose about whether the measurements supported the steady state or evolutionary models of the universe

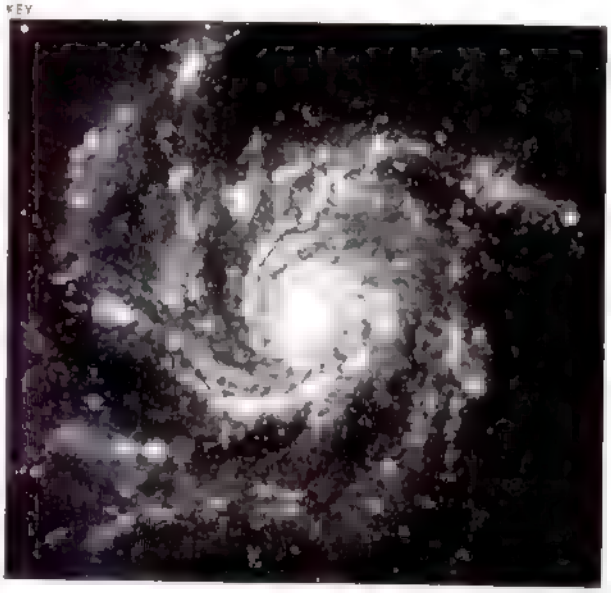
In 1965 scientists at the Bell Telephone Laboratories, New Jersey, discovered by accident a radiation from the sky with maximum intensity at a wavelength of 7cm distributed quite uniformly over the heavens. In subsequent years measurements of the spectrum of this microwave radiation seem to have confirmed their initial claim that this is the "relic" radiation from the initial high temperature-high density phase of the universe envisaged by Gamow

Questions about evolution

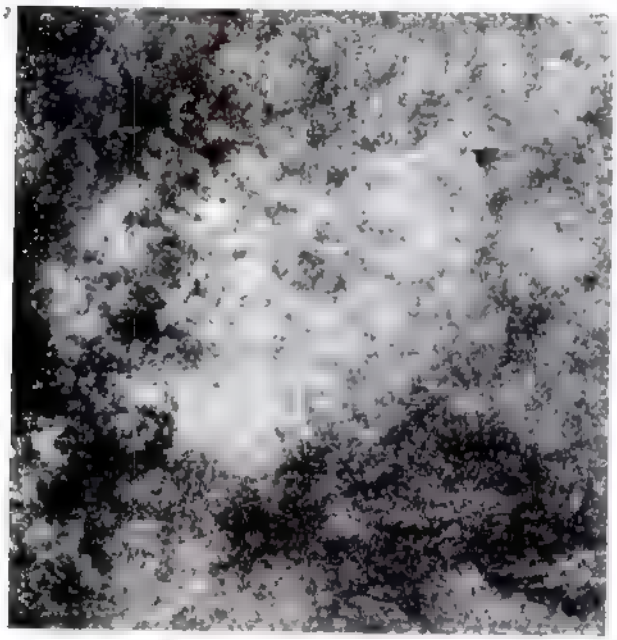
Since the discovery of the relic microwave radiation the concept that the universe is in an evolutionary, as distinct from a steady state has been widely accepted Areas of doubt persist, however One concerns the future behaviour of the universe. Is there enough material in the universe to overcome by gravitational attraction the forces of expansion? The critical density is 2×10^{-29} gm/cc but it is unlikely that measurements of this density can settle the issue because of the uncertainty about how much unobservable material exists in space. It seems more probable that the issue will be settled by observations of the remote objects in the universe such as quasars, for example, how do their recessional velocities vary with distance?

At present the results are too scattered to enable any conclusions to be made A far more pressing question concerns the state of the universe at the beginning of time. The theories propose a singular condition of infinite density at time zero. The microwave measurements probably refer to an epoch only a minute or so after the beginning of the expansion. Present physical theories can envisage a much earlier phase only a fraction of a second after the beginning of the expansion of the universe

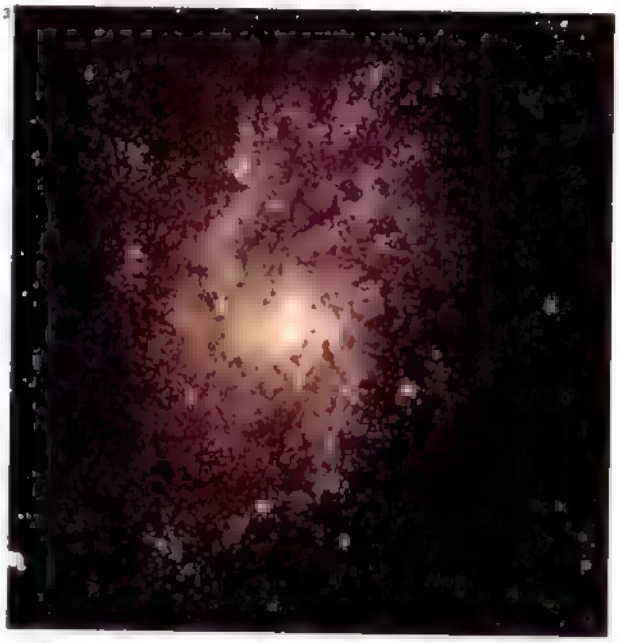
KEY



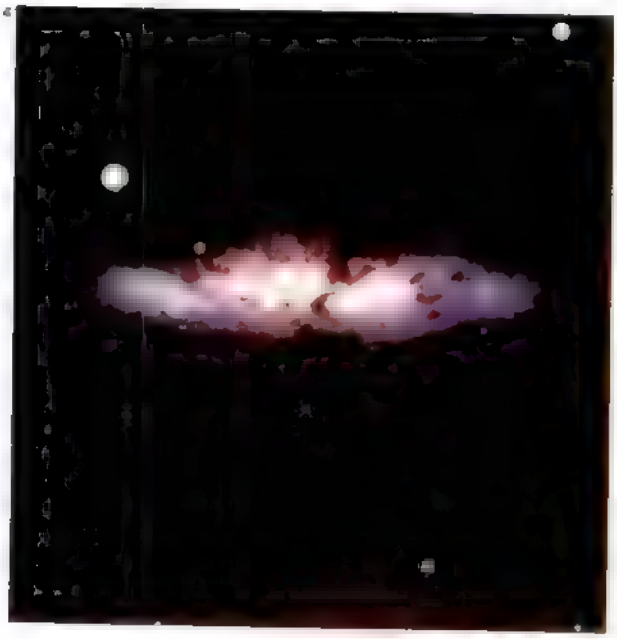
M101 in Ursa Major is a typical spiral galaxy. This is one of the systems that is close enough to be studied in detail



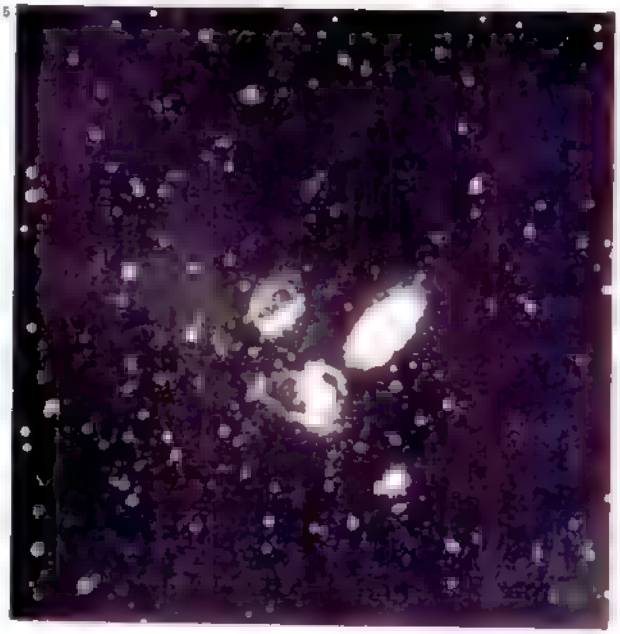
2 Star-clouds in the Galaxy, shown here in the region of Sagittarius and photographed with the Schmidt telescope, indicate the direction of the centre of our Galaxy (illustration 1). All galaxies beyond the local group are receding from our Galaxy which apart from the Andromeda system, is the largest member of the local group.



3 M33, the Triangulum Spiral, is the most distant known member of the local group of galaxies, shown in illustration 1, and is not receding from us. It has only about 1/25 the mass of our Galaxy M33 is a normal open-type spiral (Sc) The distance is 2.35 million light-years from Earth



4 M82 in Ursa Major is an irregular galaxy 10.5 million light-years away, beyond the local group Movements of the gas within it indicate that a tremendous explosion took place there 1 5 million years before our present view of it. M82 is also a strong emitter of radio waves.



5 In Stephan's Quintet of galaxies the spiral forms can be made out, but with remote systems, at the limit of observation, estimates of distance depend only upon Doppler shifts No measurements out to 10 000 million light-years, the limit shown in illustration (1), have yet been made by either optical or radio methods.

Mapping the constellations

It requires little imagination to make patterns out of the stars in the sky. There are some groups that beg for treatment: for instance the seven stars making up the so-called Plough or Big Dipper and those within the pattern of Orion. In the far south the four stars of the Southern Cross are equally distinctive, even though they look more like a kite than a cross.

At an early stage the stars were divided up into constellations, each of which was given a name. It is, however, important to remember that these visual categories are created simply by line-of-sight and that the stars in any particular constellation are not necessarily associated: thus in the Plough, the "end" star, Alkaid, is more than twice as far away from Earth as the second star, the celebrated binary Mizar. Nor are the constellations permanent: patterns change with time as the stars move.

Early star maps

Ancient star-gazers, who had no idea of the construction of the universe, believed the stars to be equally distant. If they were, the constellation patterns would presumably be of real importance. The zodiacal constellations are probably the oldest: the Babylonians traced the Zodiac and divided it into 12 constellations [1], which led to our present division of the year into 12 months. Both the Chinese [6] and the Egyptians [4] drew up maps of the sky, showing named constellations, and so, rather later, did the Greeks. It is the Greek system that has survived.

Naming the constellations

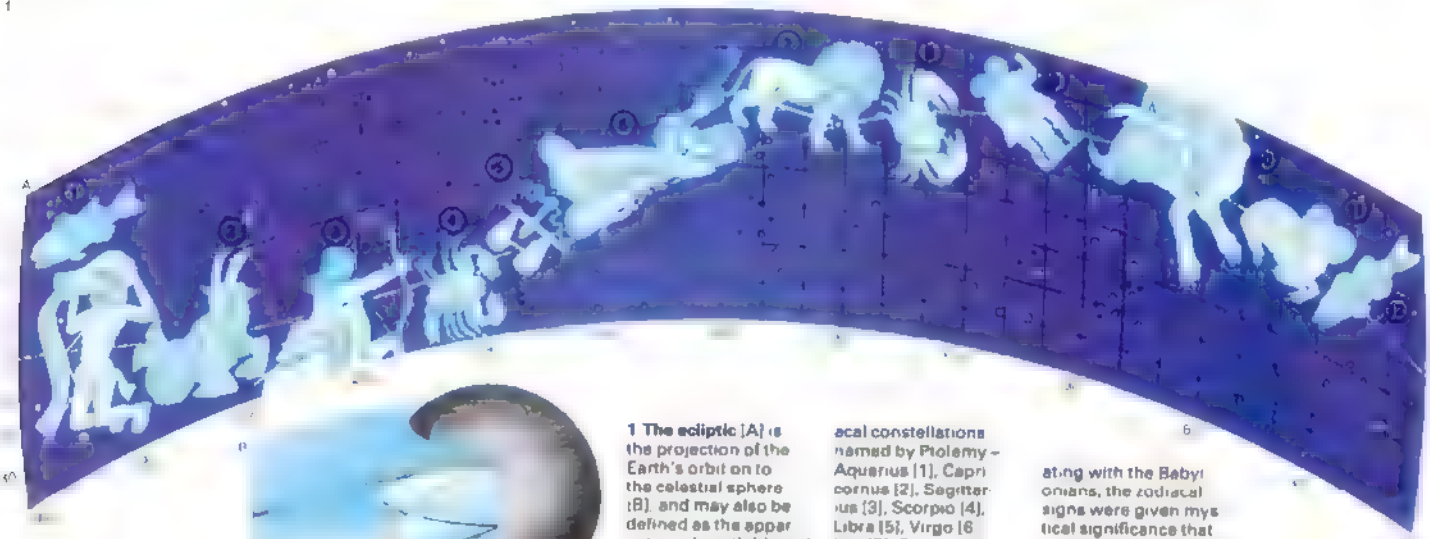
About 150 BC Hipparchus, one of the greatest astronomers of the Classical period, compiled a star catalogue. It was this catalogue that was used by Ptolemy of Alexandria as a basis for his own work, undertaken around AD 150. Ptolemy gave a list of 48 constellations, all of which are still to be found on astronomical maps even though their boundaries have been modified. Ptolemy also divided the stars up into magnitudes, or grades of apparent brilliancy, from 1 (very bright) down to 6 (the faintest stars normally visible with the naked eye on a clear night).

The "Ptolemaic" constellations were named after mythological figures, living creatures, or – a relatively few – inanimate objects. The list included virtually all the famous constellations that can be seen from the latitude of Alexandria where (as far as is known) Ptolemy spent all his life. His Latin names are still used: Ursa Major (the Great Bear), Aries (the Ram) and Aquarius (the Water-bearer), and those of mythological characters such as Perseus, Cepheus, Cassiopeia and Andromeda.

It has been claimed that the sky is a complete picture-book [1] in which the classical legends are illustrated and preserved. There is, for instance, the tale of how some tactless boasting by a proud queen, Cassiopeia, about the beauty of her daughter Andromeda led to the princess being chained to a rock on the sea-shore, there to await the coming of a monster sent by the sea-god. This is one of the legends that has a happy ending, for Andromeda was rescued by the gallant hero Perseus, who was returning from an expedition during which he had killed a hideous creature named Medusa, the Gorgon, who

CONNECTIONS

See also



1 The ecliptic [A] is the projection of the Earth's orbit on to the celestial sphere [B], and may also be defined as the apparent yearly path (dotted line) of the Sun among the stars. It passes through the 12 zodiacal constellations named by Ptolemy – Aquarius [1], Capricornus [2], Sagittarius [3], Scorpio [4], Libra [5], Virgo [6], Leo [7], Cancer [8], Gemini [9], Taurus [10], Aries [11] and Pisces [12]. Originating with the Babylonians, the zodiacal signs were given mystical significance that has continued with the modern widespread interest in astrological prediction.

2 This early Arab water clock is surrounded by an Arab zodiac, with the constellation figures clearly represented. The Arab star catalogues were more accurate than any previously made.

3 Indian constellation patterns included Makara [A], the Sea monster, and Kumbha [B], the Waterpot, containing an elixir which the gods took to heaven – an occasion celebrated by an Indian festival.

4 The impact of Greek astrology on the Egyptians can be clearly seen in these two zodiacal figures (Scorpio and Sagittarius), found inside an Egyptian mummy case, dating from the second century AD.

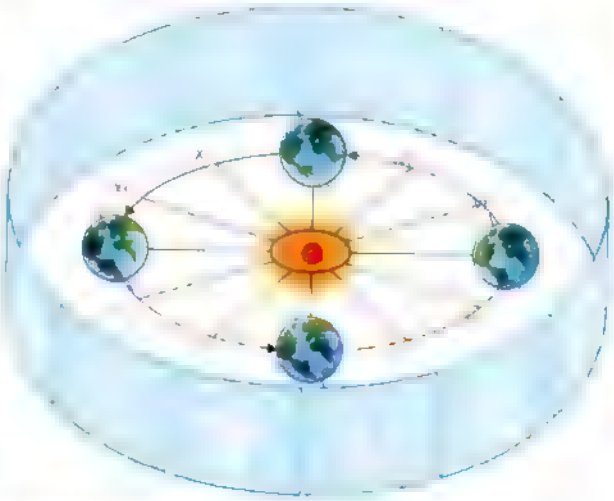
had snakes instead of hair and whose glance would turn any onlooker to stone. All the main characters in the story can be seen in the Northern Hemisphere: Andromeda, her parents Cassiopeia and Cepheus, the sea-monster (Cetus) and of course Perseus himself, with the Gorgon's head marked by the "demon star", the eclipsing binary Algol. Another legend concerns Orion, the great hunter. A scorpion was sent by Apollo who wished to protect his sister Artemis from Orion. Orion escaped from the scorpion but was killed by one of Artemis' arrows, she being unaware of whom she had shot. Artemis then set Orion's image among the stars, but far away from the Scorpion.

Extending Ptolemy's list

Few of the constellation patterns bear the slightest resemblance to the objects after which they are named; Scorpio is one of the exceptions. However, the system has been followed for so long that it will certainly never be altered now.

Ptolemy's list did not include all the stars visible from Alexandria and he could know nothing about the stars to be seen from places farther south. Later astronomers extended his list, even to the extent of stealing stars from the original 48 groups; there was a time when no astronomer seemed to feel happy without making at least one addition to an already crowded sky. The climax came with the maps by J. E. Bode, in the late eighteenth century, which included groups with barbarous names such as Sceptrum Brandenburgicum (the Sceptre of Brandenburg), Lochium Funis (the Log Line) and Officina Typographica (the Printing Press). Subsequently the list was reduced to a more manageable 88 groups and the boundaries were rigidly defined by a commission of the International Astronomical Union in 1934.

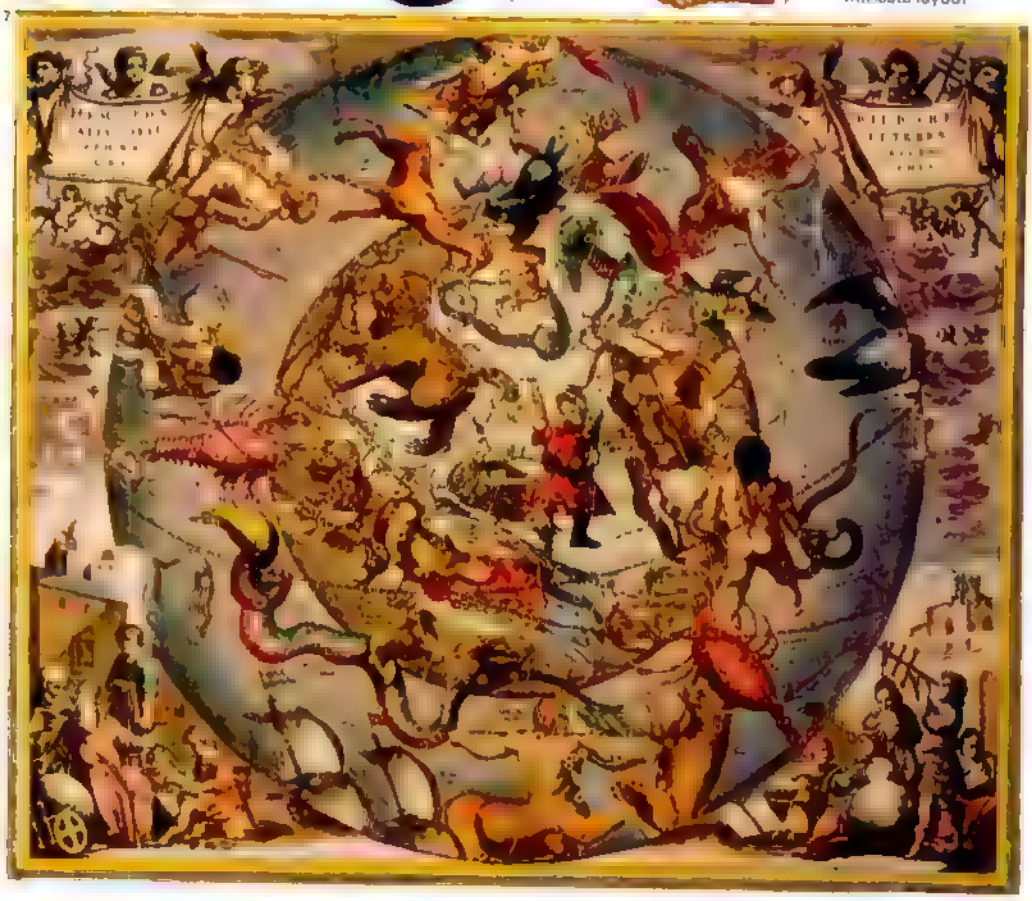
The constellations are unequal in size and in importance. One (Argo Navis, the Ship Argo) was so huge that it has been split up into several parts; the largest constellation now recognized is Hydra (the Watersnake), which sprawls its way across the sky but has only one bright star. It is interesting to note that the famous Southern Cross, Crux, is the smallest constellation; it was named in 1679.

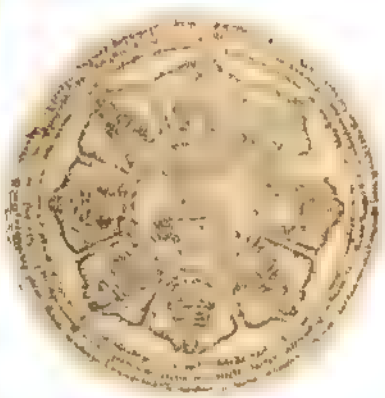


The 12 constellations of the Zodiac are shown in relation to the Sun during its apparent movement along the ecliptic – a result of the orbital motion of the Earth. An observer on Earth will see the Sun in Aries in March, but as the constellation is only above the horizon during daylight it will be unobservable.

5 Japanese zodiacal signs are animals. The Dog (A), Cock (B), Snake (C), Tiger (D), Rat (E) and Owl (F) are shown in the form of netsuke or buttons worn on a kimono.

6 Chinese zodiacal signs are also animals, some of which are depicted here on a vase (A) of the 5th or 6th century AD. This Chinese horoscope (B) has an intricate layout.





7 The main constellation patterns listed by Ptolemy are shown in this map drawn by Joannes Janssonius in 1660. The map is orientated to the pole of the ecliptic. The confusion of the intricate patterns led the 18th-century astronomer William Herschel to state that they were created to "cause as much inconvenience as possible". The patterns were subsequently modified and laid down by the International Astronomical Union in 1934. The revisions are now universally accepted.

Excited by the light of young stars
[illegible] the [illegible] shows [illegible] the [illegible] of space [illegible] a star [illegible] a huge [illegible] 1,000 light years away from us which [illegible] ionised under gravity to give birth to dozens of stars. North is to the left in this orientation.

A mighty Saturn 5 rocket lifts off from Cape Canaveral on 18 May 1969 carrying astronauts Thomas Stafford, John Young and Eugene Cernan towards the Moon. This was the final dress-rehearsal before man's first lunar landing on July 21 1969, during which every manoeuvre was practised – including a descent to within 14.5km (9 miles) of the Moon's surface itself.

Star guide: northern sky I

The far northern sky is dominated by the constellation of Ursa Major, the Great Bear, probably the most famous constellation in the sky. Its seven bright stars make up the pattern that is commonly known as the Plough or in the USA, the Big Dipper.

Ursa Major – the key constellation

The shape of the Plough is so distinctive that it cannot be overlooked. Six of the stars are of about the second magnitude, but the seventh, Delta Ursae Majoris or Megrez, is the odd third. Its relative faintness can be missed and yet instruments of recent times ranked it equal with its companions so that if their descriptions are accurate (which is by no means certain), Megrez has faded appreciably. Alpha, or Dubhe, the brighter of the Pointers to the Pole Star, is orange; the rest of the stars are white or bluish white. Mizar, or Zeta Ursae Majoris, has a fainter star, Alcor, close beside it. Through a telescope, Mizar can be seen to be a fine double.

From Britain and the northern most part of the USA, Ursa Major is a circumpolar – that is to say, it never sets – and it is extremely useful as a guide for locating other stars and constellations. Ursa Minor, the Little Bear, is easily found [Key]. It resembles a faint and distorted version of the Great Bear, but it has two stars of the second magnitude – Polaris itself and the orange Beta, or Kocab, sometimes nicknamed the Guardian of the Pole. Between the Bears there sprawls the long, dim constellation of Draco, the Dragon, whose brightest star, Gamma or Eltanin, is of about the second magnitude. Alpha Draconis, or Thuban, between Kocab and Alkaid in the Plough, used to be the pole star in ancient times.

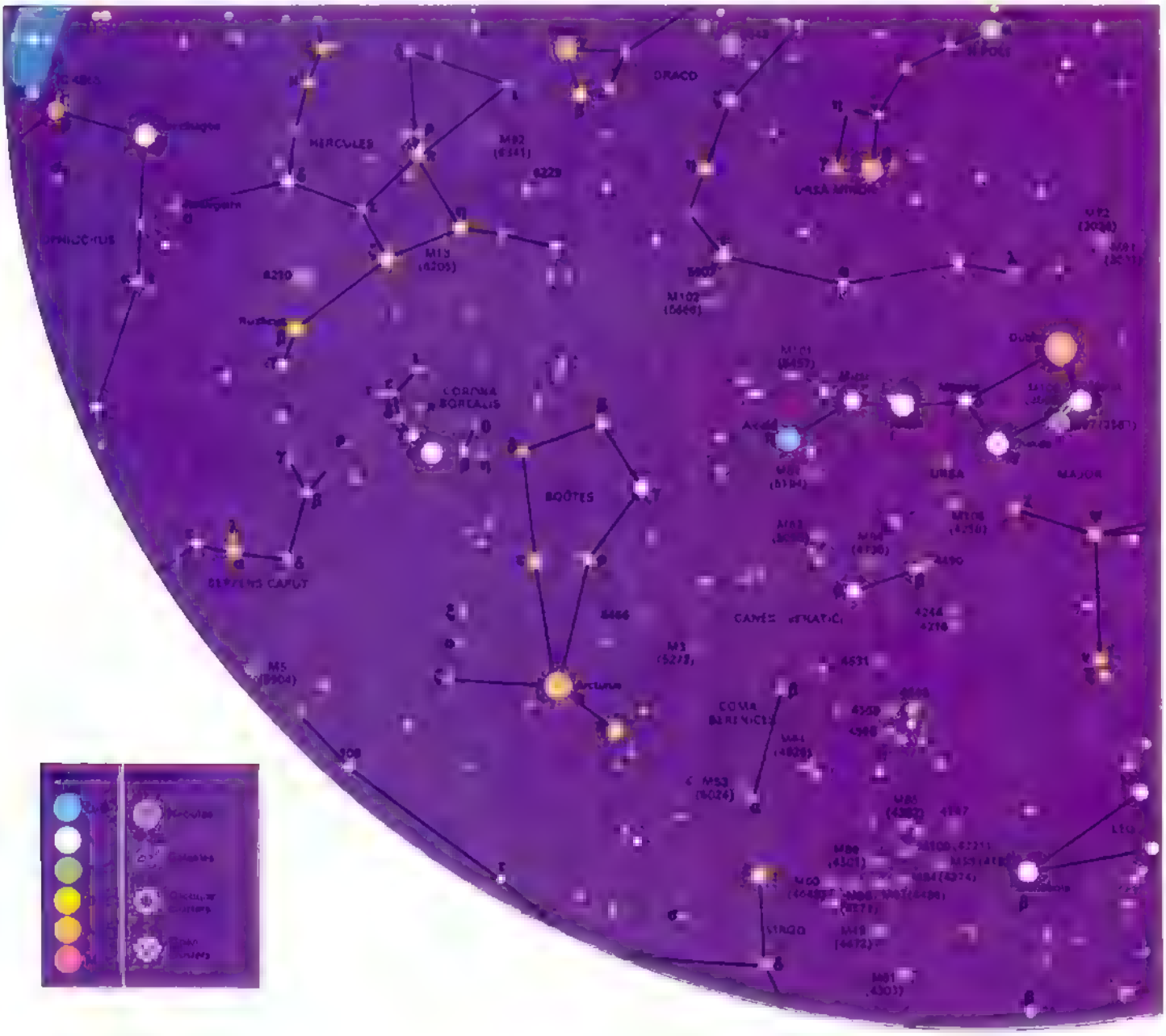
Map I – from Hercules to Virgo

In the large but rather ill formed Hercules, the brightest star, Beta, is above the third magnitude. Alpha, or Rasalgethi, is a semi-regular variable, with a range of between magnitudes 3 and 4; it is a huge red giant star with a small greenish companion visible in small telescopes. Close by it is the second-magnitude Rasalhague, the brightest star in another, larger, rather dim group, Ophiuchus, the Serpent-bearer. However, the most interesting objects in Hercules are the globular clusters M13 (NGC 6205) and M92 (NGC 6341). M13, the finer, is just visible with the naked eye and a fairly small telescope will resolve its outer portions into stars.

Following round the "tail" of the Great Bear will lead to Arcturus, the brilliant orange star in Bootes, the Herdsman. Arcturus is the brightest star in the northern hemisphere of the sky. Its magnitude is 0.06 and it is light orange in colour, with a K-type spectrum. It is 36 light-years away from Earth and 100 times more luminous than the Sun. The rest of Bootes is not notable, although Epsilon (Izar) is a beautiful double star. Close to Bootes is the conspicuous little semicircle of stars making up Corona Borealis, the Northern Crown, which contains the celebrated variable star R Coronae. This is normally of about the sixth magnitude, but suffers sudden, unpredictable drops to minimum. R Coronae is the prototype star of its class and is much the brightest example. Also in this constellation is T Coronae, which is normally of the tenth magnitude, but flared up to naked-eye visibility

CONNECTIONS

See also



in 1866 and once again 80 years later.

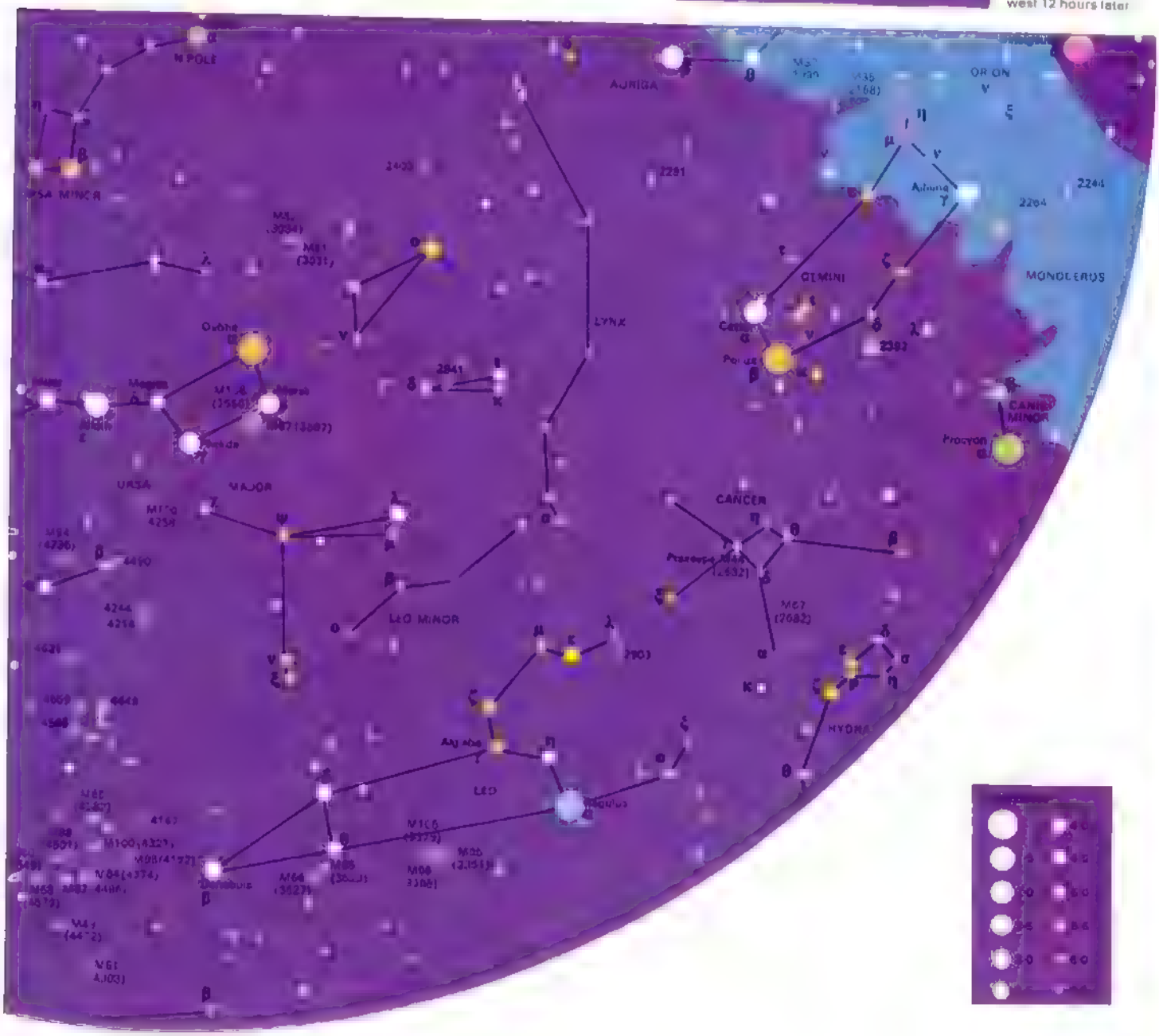
Part of Virgo is also shown, although its leading star, Spica, is in the southern hemisphere. This region is particularly rich in faint galaxies and between Virgo and Ursa Major lies the constellation of Coma Berenices (Berenice's Hair), which looks almost like a large, dim cluster. Canes Venatici, the Hunting Dogs, has only one star as bright as the second magnitude.

Map 2 – from Leo to Canis Minor

Leo, the Lion, is the most easily visible of the constellations during spring in the Northern Hemisphere. Its leading star, Regulus, lies at one end of a curved line making up the pattern known as the Sickle and is of magnitude 1. Gamma, or Algieba, is a fine double star with rather unequal components. Beta or Denebola, on the other side of Leo, is now of the second magnitude, but in ancient times it was ranked as being of the first and there is a chance that it, too, has faded appreciably. Adjoining Leo is Cancer, the Crab, which contains the famous open clusters M44 (NGC 2632) or Praesepe, easily visible with the naked eye on a dark night, and M67 (NGC 2692), which is visible with binoculars and is thought to be one of the oldest of the loose star clusters.

The brilliant constellation of Orion is cut in half by the celestial equator, so that part of it is shown on this map; the leader is the orange-red Betelgeuse. Not far from Orion are the Twins, Castor and Pollux, in the constellation of Gemini. Pollux is of the first magnitude and Castor between the first and the second – another possible case of fading. Pollux has a K-type spectrum and is clearly orange. Castor is a multiple star, made up of two main components, each of which is a very close binary, together with a fainter companion that is also a binary. The other bright star in Gemini is Gamma or Alhena (magnitude 2). Adjoining Gemini is Canis Minor, the Little Dog, with one brilliant first magnitude star, Procyon, which is one of our nearest stellar neighbours and has a white dwarf companion. The Milky Way flows through Gemini and its neighbour Auriga, the Charioteer, resulting in a region of many rich star-fields.

The north celestial pole is easy to find because Polaris lies within one degree of it. Polaris can be found by following the line of the Pointers, Merak and Dubhe, in the Great Bear. As the Earth rotates from west to east, the celestial pole remains stationary while all other objects appear to circle it slowly. Depending on the position of the observer, the Pole Star will be in a different area of the sky, only at the North Pole is it seen directly overhead. As the celestial pole is thus an extension of the Earth's axis, so also is the celestial equator a projection of the Earth's Equator. A star at the celestial equator will rise in the east and set in the west 12 hours later.



Star guide: northern sky II

Cassiopeia, whose leading stars make up a well-marked W or M formation, is second in importance only to Ursa Major among the constellations of the far north

Cassiopeia – the key constellation

Like Ursa Major, Cassiopeia is circumpolar over Britain and the northern USA. The best way to find it is to extend a line from Mizar the second star in the tail of the Great Bear through Polaris and along a line for an equal distance in the opposite direction. Gamma Cassiopeiae, the middle star of the W, is an irregular variable. It is usually just below the second magnitude, but sometimes, as it did in 1936, it can flare up to 1.6. Its spectrum is peculiar and it is a highly unstable star. Alpha or Shedir is of type K and is suspected of slight variability. Beta is invariable at magnitude 2.3. Tycho Brahe's supernova of 1572 flared out in Cassiopeia; it is still a distinctive source of radio signals.

Two of the stars in the W point to Perseus which has one second-magnitude star, Alpha or Mirphak, as well as the celebrated eclipsing binary Algol (Beta) which has a magnitude range of 2.2 to 3.5. Perhaps the most notable object in Perseus is the Sword Handle, Chi-h (known officially as H VI 33–4), which consists of two beautiful open clusters in the same telescopic field, making up a glorious spectacle. Each cluster has a diameter of 75 light-years and lies at a distance of 7,000 light-years from Earth.

Map 1 – from Auriga to Triangulum

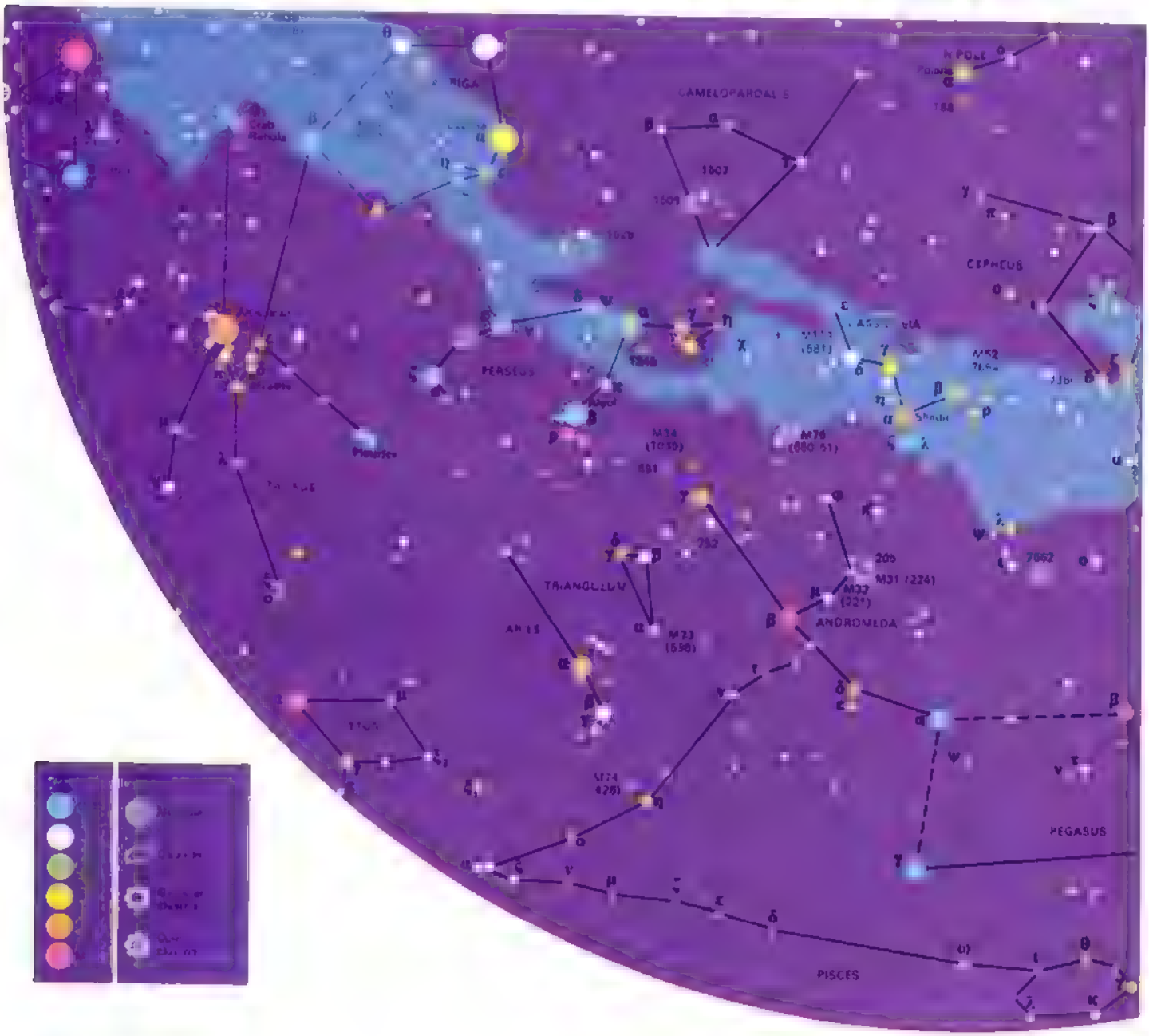
The Milky Way flows through Cassiopeia and Perseus and on into Auriga, where the leading star is Capella – almost exactly equal in brilliancy to Vega and on the opposite side of the celestial pole. From Britain it is almost overhead during winter evenings. Capella is 45 light-years away and is of the same spectral type as the Sun, although it is a giant star and much more luminous. The small triangle beside it makes up the Haedi, or Kids. Epsilon Aurigae, the apex of the triangle, is a remarkable object; it is a binary and the secondary is either a young star or a black hole. Zeta Aurigae close beside it, is also an eclipsing binary of long period (972 days). Auriga itself is made up of a quadrilateral of stars easy to identify; the constellation includes several bright, open clusters.

Taurus, the Bull, adjoins Auriga. Aldebaran, orange and of the first magnitude, lies in line with the three stars of Orion's belt, which are just in the southern hemisphere. Taurus contains the two most famous open clusters in the sky, the Pleiades and the Hyades, as well as the Crab Nebula, M1, near the third-magnitude Zeta Tauri.

Cassiopeia can also be used to locate the Square of Pegasus, which is prominent during autumn evenings in the Northern Hemisphere. Of the four leaders, Beta, or Scheat, is semi-regular with a period of approximately 35 days; it is visible as a huge, red giant. The line of stars leading off from Pegasus make up Andromeda – most celebrated because of the presence of the spiral galaxy M31. Alpheratz or Alpha Andromedae, is included in the Square of Pegasus. Of the other second-magnitude leaders of Andromeda, Beta is [illegible] and Gamma [illegible] is [illegible] Lovely clusters [illegible] well [illegible] telescope. Pegasus contains [illegible] bright star, Enif, [illegible] from the Square. Epsilon [illegible] Adjoining

CONNECTIONS

See also



the Square is the dim zodiacal constellation of Pisces, and in this region also lie Aries, with the second-magnitude Hamal (Alpha) and Triangulum. Triangulum contains the loose spiral galaxy M33, a member of our local group of galaxies, and visible with binoculars on a clear night.

Map 2 – from Lyra to Delphinus

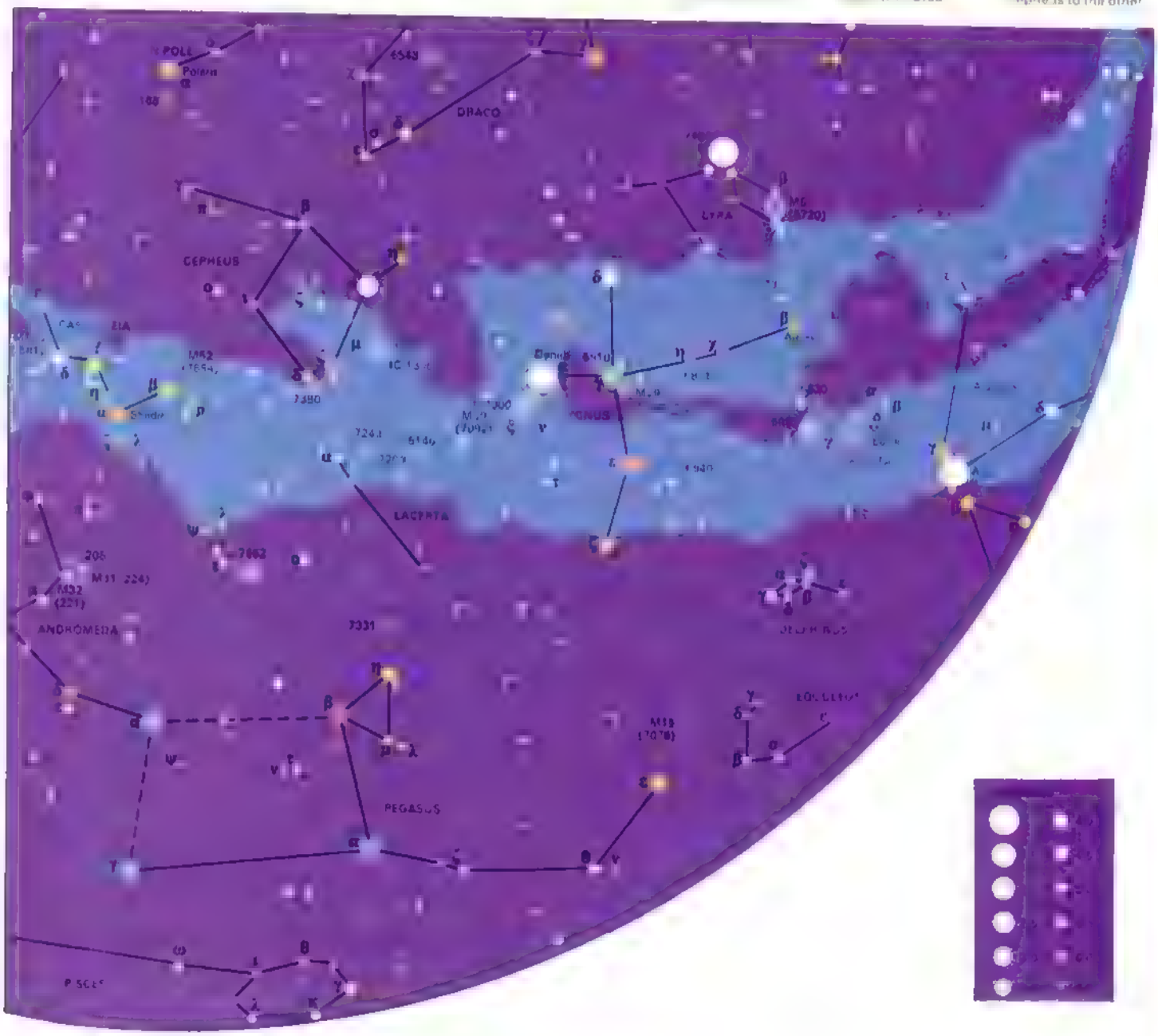
Vega, in Lyra, is one of three brilliant stars making up what has been unofficially called the Summer Triangle. Vega is the fifth brightest star in the whole sky and is easy to identify. From Britain and the northern United States it is almost overhead during summer evenings. Lyra is a small constellation, but contains a great many interesting objects. Epsilon Lyrae is a quadruple star, while Beta is the famous eclipsing variable. M57 (NGC 6720) is the most famous of the planetary nebulae, which are surrounded by gaseous shells, and lies between Beta Lyrae and the third magnitude Gamma.

The other two members of the Summer Triangle are Deneb in Cygnus and Altair in Aquila. Cygnus is a superb group. It is shaped rather like an X and among its many interesting objects are the beautiful double Beta or Albireo, in which the primary is golden-yellow (magnitude 5) and the companion blue (magnitude 3), and the long-period variable Chi, between Albireo and the centre of the X, which has a great magnitude range (3.3–14.2) and a period of 407 days. Like most of its kind it is red and at maximum is easy to locate. The Milky Way is particularly rich in this area, so that Cygnus is well worth looking at through binoculars.

Altair, in Aquila, is recognizable, partly because of its brightness and partly because it has a fainter star to either side of it. Close to it is a line of stars, of which the central member, Eta Aquilae, is a typical Cepheid, with a period of 7.17 days.

In the general area of Cygnus and Aquila there are some small but quite distinctive constellations, notably Delphinus, the Dolphin. It was here that the famous slow nova HR Delphini, appeared in 1967, it rose to magnitude 3.6 and is still visible with a small telescope. Sagitta is another small but easily identified group.

The area of the North Pole is marked by Polaris, which is within one degree of the pole itself. With Ursa Ma[illegible] and the rather form[illegible] asterism is to the other



Star guide: southern sky I

There can be little doubt that the stars of the far south are more splendid than those of the far north. Brilliant constellations such as Centaurus, Carina and above all Crux are not visible from Europe or most of the United States, and the same is true of the bright external systems known as the Nubeculae or Magellanic Clouds. Other stars, such as Sirius, which are visible from the Northern Hemisphere, are much brighter seen from south of the Equator.

Crux – the key constellation

Crux, the Southern Cross, is the most famous of all the far southern groups and is, incidentally, the smallest of all the recognized constellations. Since it is not at all visible from the Northern Hemisphere it is not one of the ancient constellations, and it was not added to star maps until the seventeenth century. It is not like a cross because there is no central star to make up the X as there is with Cygnus in the northern sky. Crux more nearly resembles a kite but it is so compact that it cannot be mistaken. Acrux or Alpha Crucis, the leader, is a fine binary 270 light-years from Earth. The magnitudes of the components are 1.6 and 2.1, giving a combined naked-eye magnitude of 0.8. The pair may be seen well with a small telescope. Beta Crucis (magnitude 1.3) is a very luminous B-type star and Gamma Crucis (magnitude 1.6) is a red giant. The fourth star of the pattern is much fainter at magnitude 3 and thus spoils the symmetry. Crux also includes the Jewel Box cluster, Kappa Crucis (NGC 4755), and the dark nebula that is usually termed the Coal-sack – a huge mass of dust and gas cutting out the light of stars beyond. Most astronomers consider this to be the finest of the dark nebulae.

Map 1 – from Carina to Hydra

Carina is part of the old constellation of Argo Navis, the ship that carried Jason and his companions in search of the Golden Fleece – but Argo was so large that it was cut up into a Keel, a Poop and Sails (Carina, Puppis and Vela). Carina contains many bright stars including Canopus, an F-type supergiant of magnitude –0.7, and also the extraordinary object Eta Carinae, which is wreathed in nebulosity and is variable. For a period between 1834 and 1844 it ranked as one of the brightest stars in the sky at magnitude –0.7 but for almost a century now it has been below naked-eye visibility at magnitude 7.7.

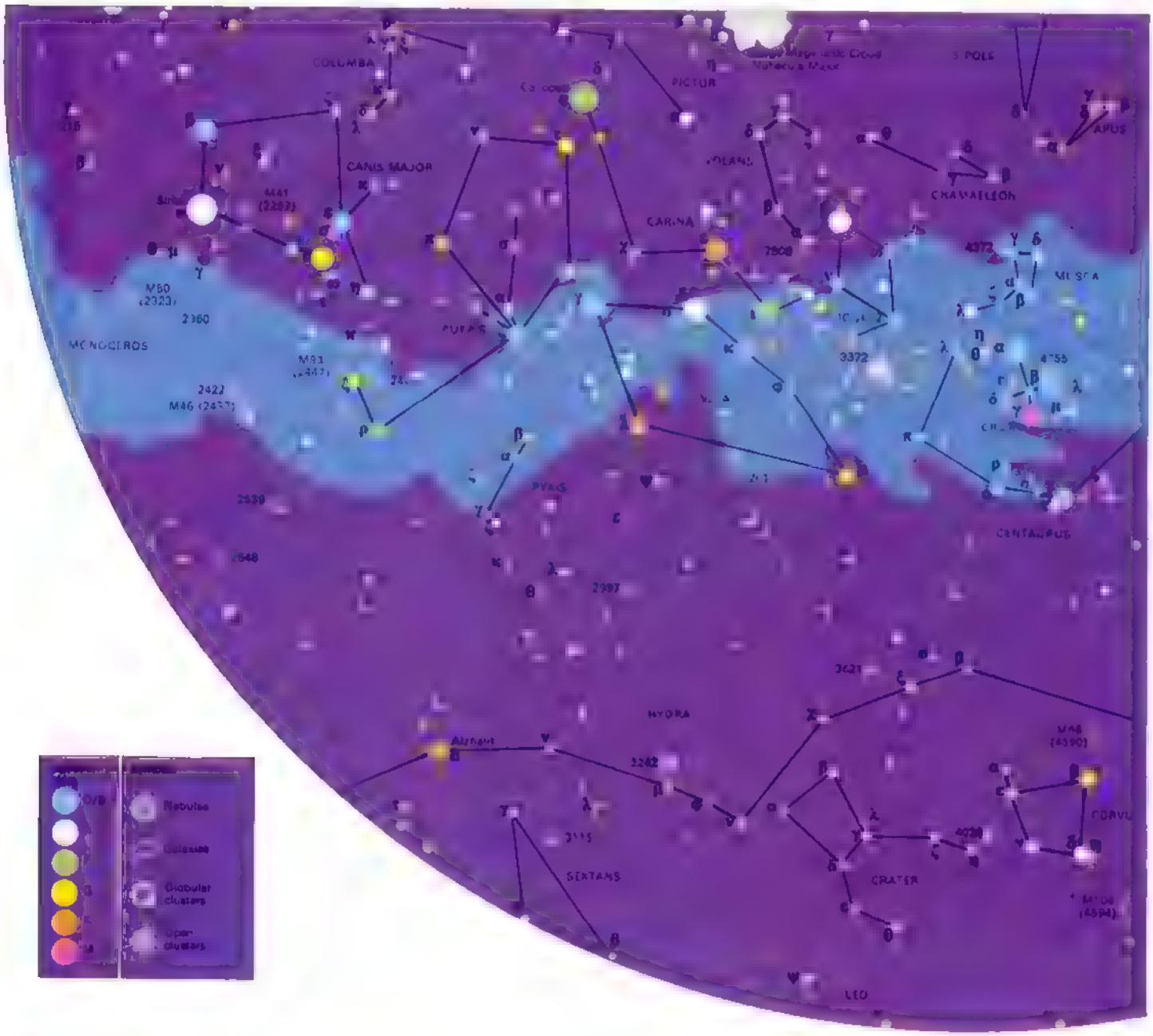
The so-called False Cross is made up of two stars in Carina and two in Vela, all of about the second magnitude. The Milky Way flows through the Ship and the region abounds with clusters and rich star fields. By contrast Canis Major is relatively barren of interesting telescopic objects, but it contains several bright stars as well as Sirius. Although it is at least 20 times more luminous than the Sun, Sirius owes its apparent eminence to its closeness (8.7 light-years from Earth) rather than to its power. It is feeble compared with Canopus which appears appreciably fainter (magnitude –0.7 as against –1.47), but is much farther away and extremely luminous.

Map 2 – from Hydra to Scorpio

Hydra, the Watersnake, is a barren area of the sky. However, it does contain the reddish second-magnitude star Alphard, the solitary

CONNECTIONS

See also



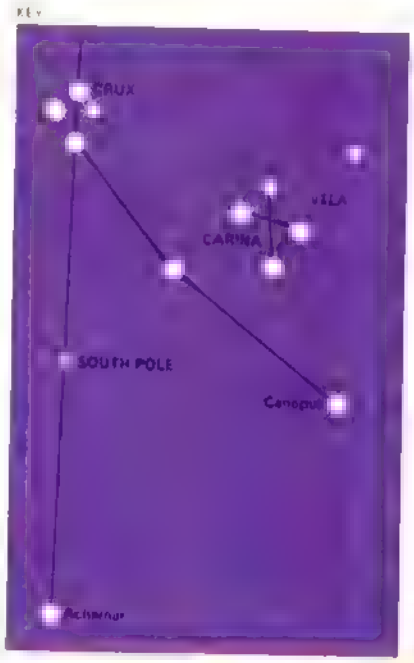
one, which appears distinct against its isolated background. The conspicuous quadrilateral of Corvus, the Crow, is also shown, and part of Virgo, including the first-magnitude Spica. Gamma Virginis, or Arich, is a fine binary with a period of 180 years.

Crux is more or less surrounded by Centaurus which is yet another magnificent group with many brilliant stars. Alpha Centauri, sometimes called Rigil Kent or Toliman, is a binary with a period of 80 years; any small telescope will separate its components of types G and K. At its distance of 4.3 light-years it is the nearest of all the bright stars, and its faint red dwarf companion Proxima, the nearest star to Earth, is only slightly closer. Adjoining Alpha is the remote Beta Centauri, at magnitude −4.3; it is more than 4,000 times as luminous as the Sun and bluish-white in colour. Also in Centaurus lies Omega (NGC 5139), the finest of all the globular clusters – conspicuous to the naked eye as a hazy patch and resolvable with a small telescope.

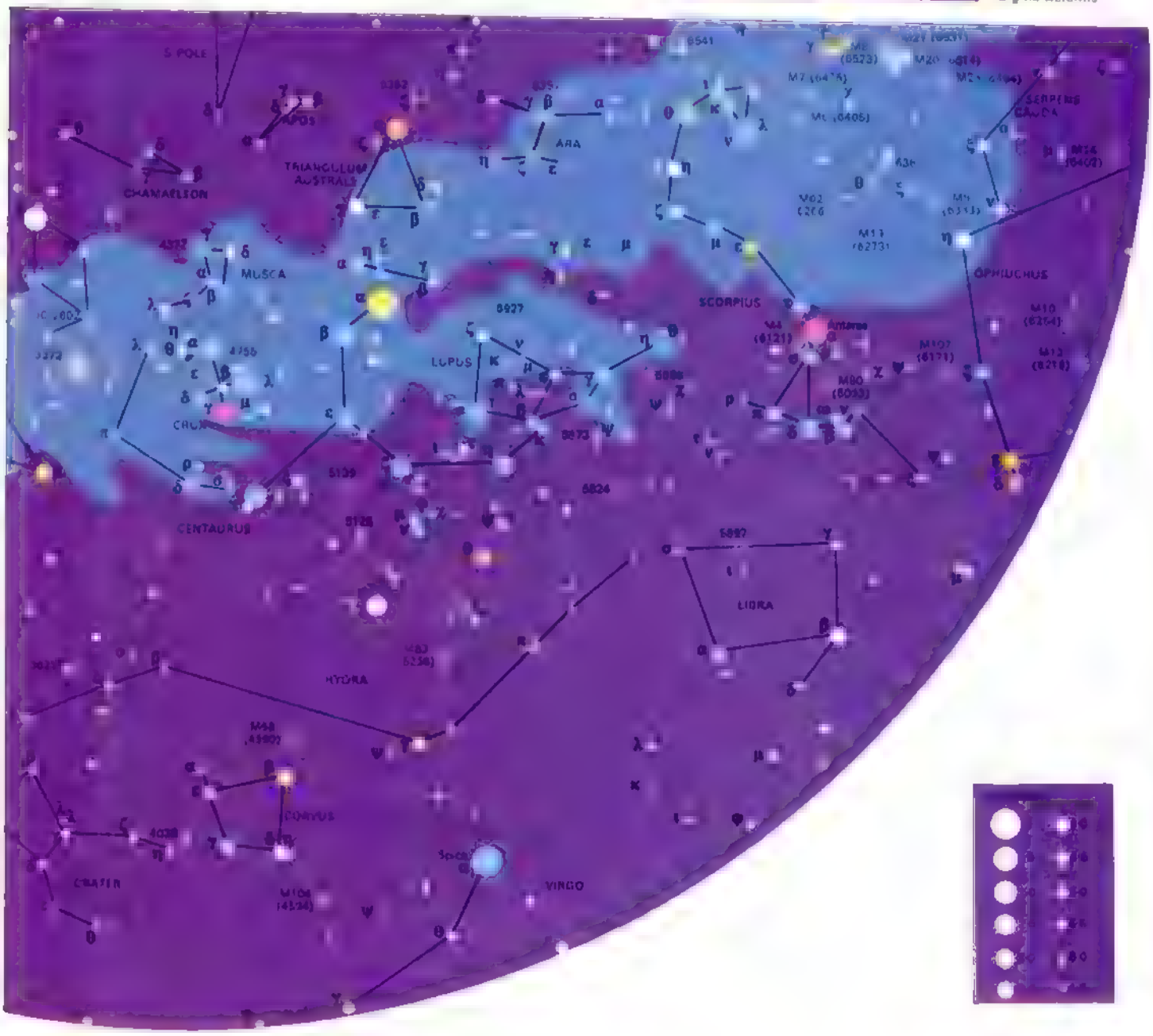
Adjoining Centaurus are Lupus, the Wolf, and Triangulum Australe, the Southern Triangle. Lupus is rather formless, but the Triangle is distinctive. Its leader, Alpha, is strongly orange in hue and is of magnitude 1.9. The other two members of the Triangle, Beta and Gamma, are of the third-magnitude.

Also shown on this map is Scorpio or Scorpius (the Scorpion), which is one of the most distinctive of the zodiacal constellations and one of the few which slightly resembles its namesake, as it consists of a long line of stars, many of them bright. Its brightest star, Antares, is a vast red giant, with a diameter of about 420 million kilometres (260 million miles); it is about 400 light-years away and its luminosity is almost 5,000 times that of the Sun. It has an apparent magnitude of 1, and rises well above the horizon over most parts of Europe.

Next to Scorpius is the obscure zodiacal constellation of Libra. The brightest star, Beta Librae, is of magnitude 2.7. Delta Librae is an eclipsing binary, with a magnitude range from 4.8 to 5.9. It is of Algol type and bright enough to be seen throughout its period with a pair of binoculars.



The south celestial pole is, compared to its counterpart in the north, more difficult to locate. The north celestial pole is conveniently marked by a bright star, Polaris, in the constellation of Ursa Minor, which lies within one degree of the polar point. Unfortunately there is no convenient south polar star; the pole lies in a barren region made up of the faint constellation of Octans. The best way to locate the pole is to follow the longer axis of Crux to the point about midway between Crux and the bright star Achernar, in the constellation of Eridanus. The nearest naked eye star to the pole is the fifth-magnitude Sigma Octantis.



Star guide: southern sky II

Because of the great quantities of dust that lie in the main plane of the Galaxy, it is impossible to see through to the galactic centre, but at least its position is known. It lies beyond the star clouds of Sagittarius, at a distance of about 33 000 light-years from Earth.

Sagittarius – the key constellation

The constellation of Sagittarius (the Archer) is not hard to identify: it contains several reasonably bright stars, although there are no [illegible]. It is, however, the southernmost constellation of the Zodiac and so is never well seen from Europe. In shape Sagittarius is rather hard to define: some people with vivid imaginations have compared it to a teapot. It adjoins the Sting of Scorpius and between the Sting and the leader of Sagittarius (Epsilon Sagittarii, or Kaus Australis) there are two bright open clusters, M6 and M7. Sagittarius is also rich in globular clusters.

Strangely enough, the star catalogued as Alpha Sagittarii is obscure: the Greek letters are not necessarily in sequence. Near Alpha is the circle of stars marking Corona Australis, the Southern Crown, which is not nearly as prominent as the Northern Crown but easier to identify.

If Sagittarius is an especially rich region, the south pole of the sky is particularly barren. There is no bright south polar star; the nearest naked-eye object is Sigma Octantis, which is a star of only the fifth magnitude and a rather poor substitute for the northern Polaris.

Map 1 – from Grus to Capricornus

The four "Southern Birds" are shown: Grus (the Crane), Pavo (the Peacock), Tucana (the Toucan) and Phoenix. This is admittedly a confusing area because only Grus has a distinctive form; it really does give the impression of a flying crane. The leaders, Alpha (Alnair, magnitude 2.11) and Beta (magnitude 2.2) are quite different. Alnair is bluish-white, but Beta is orange, and the difference is striking when the stars are seen through binoculars or any telescope. Adjoining Grus is Piscis Austrinus, the Southern Fish, which has as its leader the first-magnitude Fomalhaut – 23 light-years distant and about 15 times as luminous as the Sun. Tucana is the most obscure of the four birds, but it contains a fine double, Beta, and 47 Tucanae, a fine globular cluster second only to Omega Centauri.

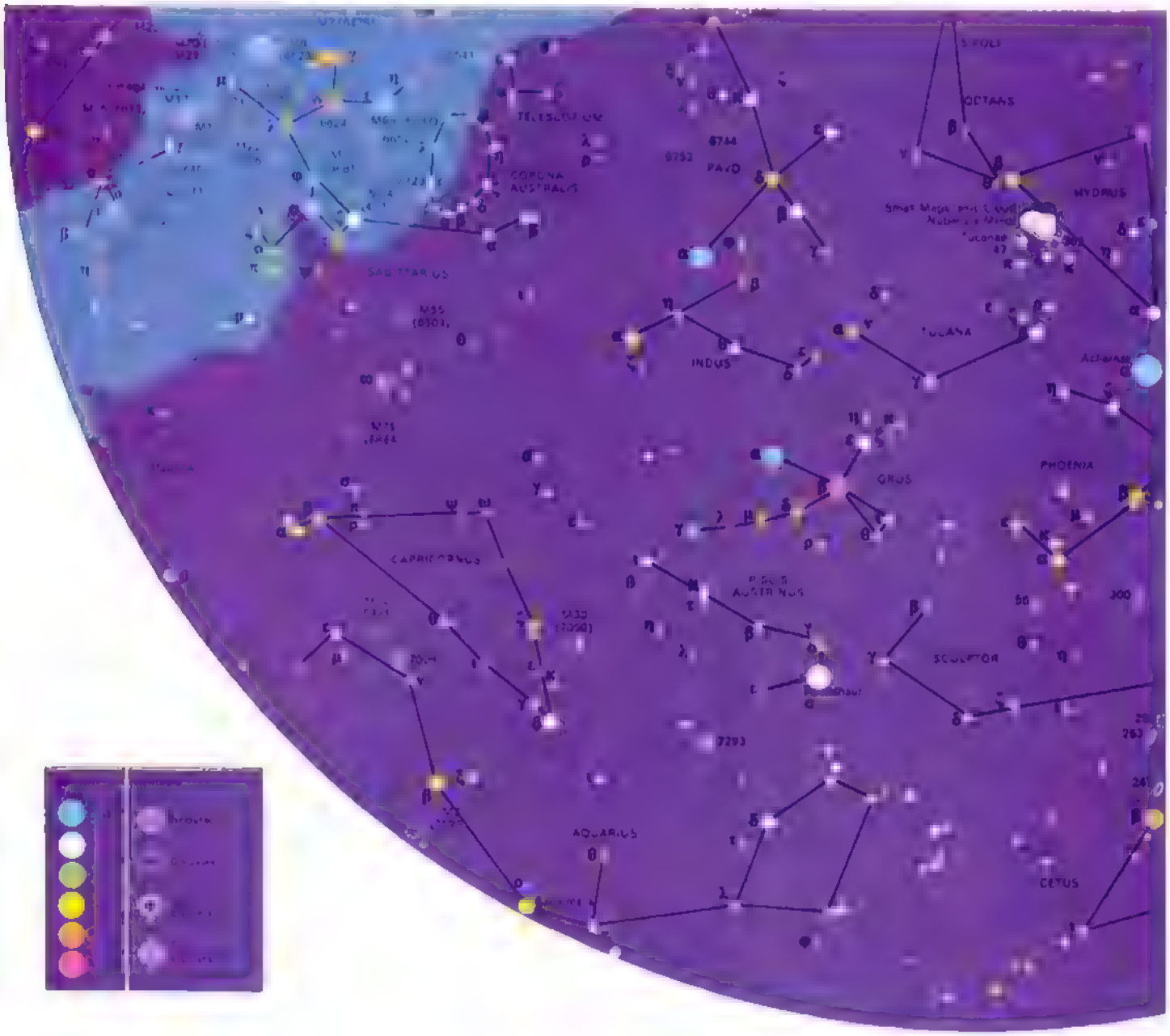
Finally, there are two obscure constellations of the Zodiac – Aquarius and Capricornus. There are a few interesting objects in Aquarius, notably the bright globular cluster M2, while in Capricornus there is the naked-eye double Alpha or Al Giedi. Beta Capricorni is also a wide, easily seen double, readily separable with a small telescope, or even with binoculars.

Map 2 – from Cetus to Orion

Cetus, the Whale or Sea-monster, is a long, rather faint constellation, most of which lies in the southern hemisphere, although the head is just north of the equator. It has one second magnitude star, Beta or Diphda, which is suspected of variability. Here too is Omicron Ceti or Mira, the most celebrated long-period variable in the sky. It has a mean period of 331.6 days and at some maxima it has been known to exceed the second mag

CONNECTIONS

See also

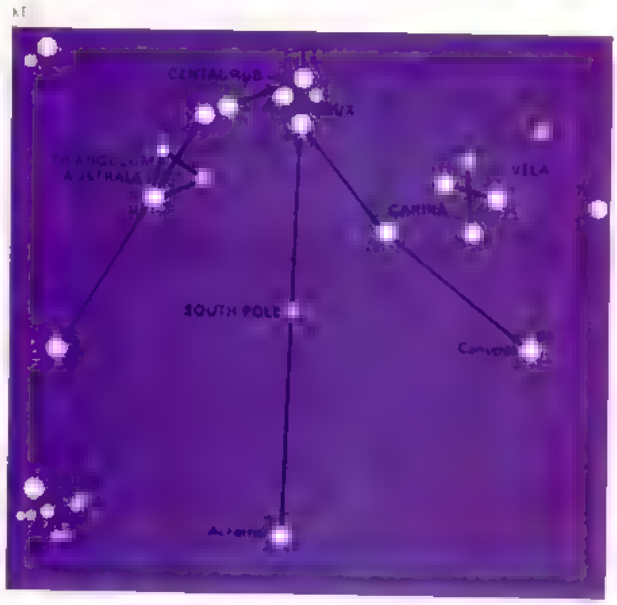


nitude, at minimum it descends to the tenth. Its variability was recognized as long ago as 1638. However, for much of the year it is below naked-eye visibility. Mira is a red giant and its colour is extremely prominent.

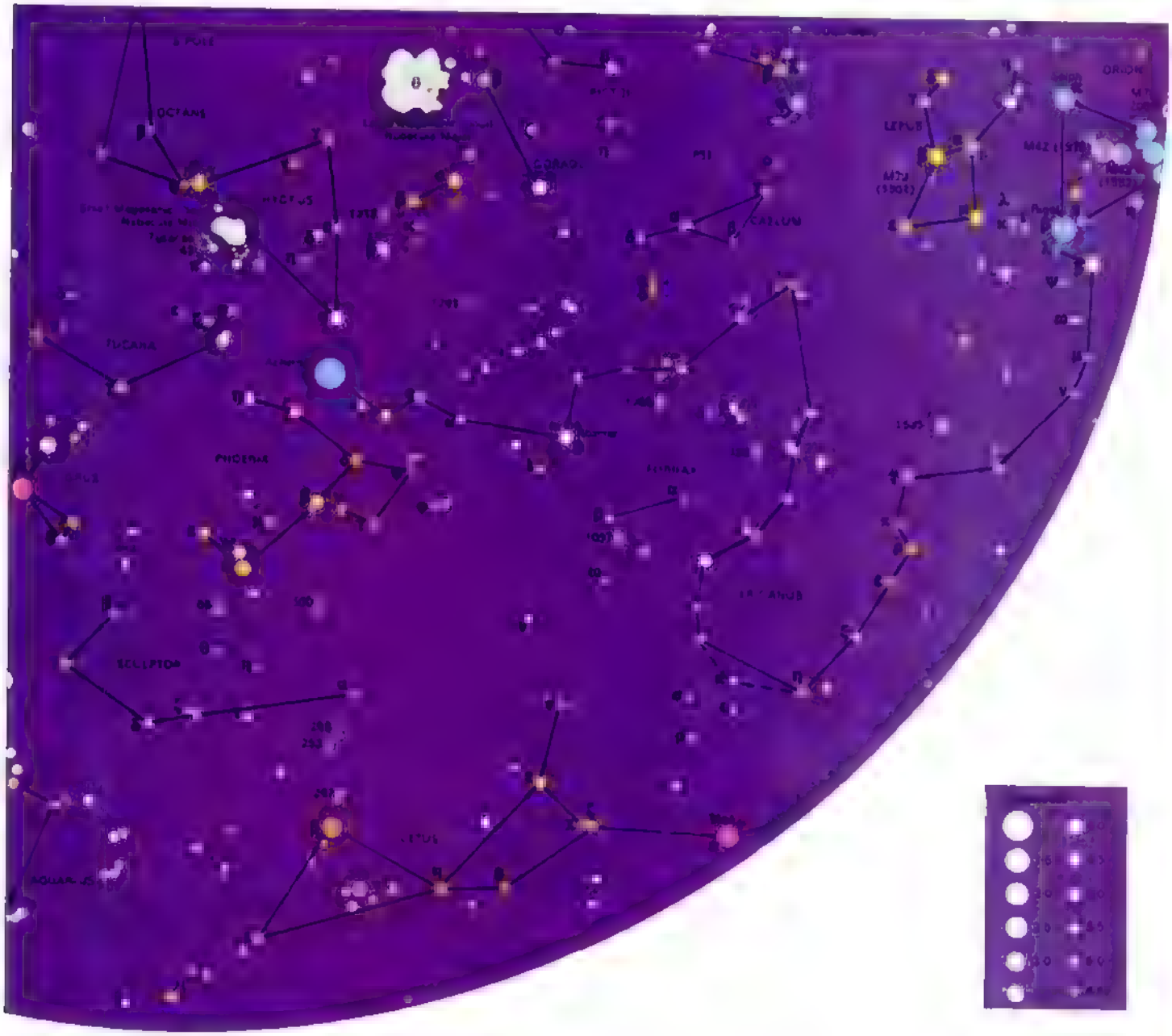
Close to Phoenix lies Achernar, the "End of the River" and the only really bright star in the long constellation of Eridanus, which winds its way from near the south pole as far as the boundaries of Orion. Achernar (magnitude −1.3) is 75 light-years from Earth and has a luminosity 256 times that of the Sun. Farther along the line of Eridanus is Theta, or Acamar, which is a fine double, and – like Castor in Gemini and Megrez in the Great Bear – is suspected of having faded during historical times, because ancient astronomers ranked it of the first-magnitude and it is now below the third.

Not far from the south pole the two remarkable Clouds of Magellan can be seen. Unfortunately they can never be seen by observers in Europe or the United States. The clouds are external systems, even though they look at first glance like detached portions of the Milky Way. They are about 150,000 light-years away from the Earth and are therefore the most remote objects clearly visible with the naked eye apart from the spiral galaxy M31 in Andromeda. The Large Cloud is so bright that even moonlight will not hide it. Binoculars bring out its form well and telescopic research has shown that it contains objects of all kinds, including globular clusters and gaseous nebulae. One star, S Doradûs, is thought to be a million times more luminous than the Sun, and yet without optical aid it cannot be seen. The Large Cloud also contains the Great Looped Nebula which is visible to the naked eye.

Orion, divided by the equator of the sky, is visible from every inhabited part of the world. The equator passes near Delta or Mintaka in the Belt, so that the brilliant Rigel lies well in the southern hemisphere. Rigel (magnitude −7.0) is a highly luminous star, thought to be about 49,000 times as powerful as the Sun and lies at a distance of 850 light-years from Earth. Also in the southern part of Orion lies the Great Nebula, M42, which contains the multiple star Theta Orionis, often known as the Trapezium.



The south polar area contains no bright star, but it is surrounded by brilliant constellations such as Crux and Centaurus.



Seasonal star maps: northern

All the stars of the northern sky are visible to an observer in the Northern Hemisphere in the course of a year. The charts on these two pages are suitable for observers living between latitudes 30° and 50°N (Key). The horizon is given by the latitude marks near the bottom of the charts. Thus for an observer who lives at latitude 30°N the northern horizon on chart 1 [N] will pass just above Deneb and Deneb will be invisible.

Limits of visibility

A star rises earlier, on average, by two hours each month; thus the chart for 22.00hrs on 1 January will be the same as that for 20.00hrs on 1 February, 18.00hrs on 1 March and so on. All the times given in the charts are in Greenwich Mean Time with no allowance made for adjustments such as British Summer Time.

A star that never sets is said to be circumpolar; thus Ursa Major is circumpolar from England while Arcturus is not. Limits on the visibility of a star for an observer at any latitude can be worked out from its declination (that is, the star's angular distance north or south of the celestial equator). To an observer in the Northern Hemisphere, a star is at its lowest point when it is below the pole and therefore due north. The circumpolar region of the sky can be found by subtracting the observer's latitude from 90°. Suppose, for example, that an observer lives at latitude 51°N. Subtracting 51 from 90 gives 39: a star which is north of declination +39° will never set, while a star south of declination −39° will never rise. Thus, from latitude 51°N, which is approximately the latitude of London, the star Capella (declination +46, in round figures) will be circumpolar while Arcturus (declination +19) will be well seen but will not be circumpolar. Canopus (declination −53) will remain permanently below the horizon from this latitude.

Alkaid or Eta Ursae Majoris, the most southerly of the seven principal stars of the Great Bear, has a declination of approximately +50°. It will therefore be circumpolar from latitudes north of 40°N (90 − 50 = 40) and it will be invisible to an observer from latitudes south of 40°S.

The charts show the southern [S] and the northern [N] aspects of the sky from the viewpoint of an observer in northern latitudes; the descriptions that follow are for the late evening (when the sky appears the same as it did three months previously). Inevitably, charts of this kind involve some distortion, but used solely as recognition aids they are quite satisfactory.

Stars of chart 1

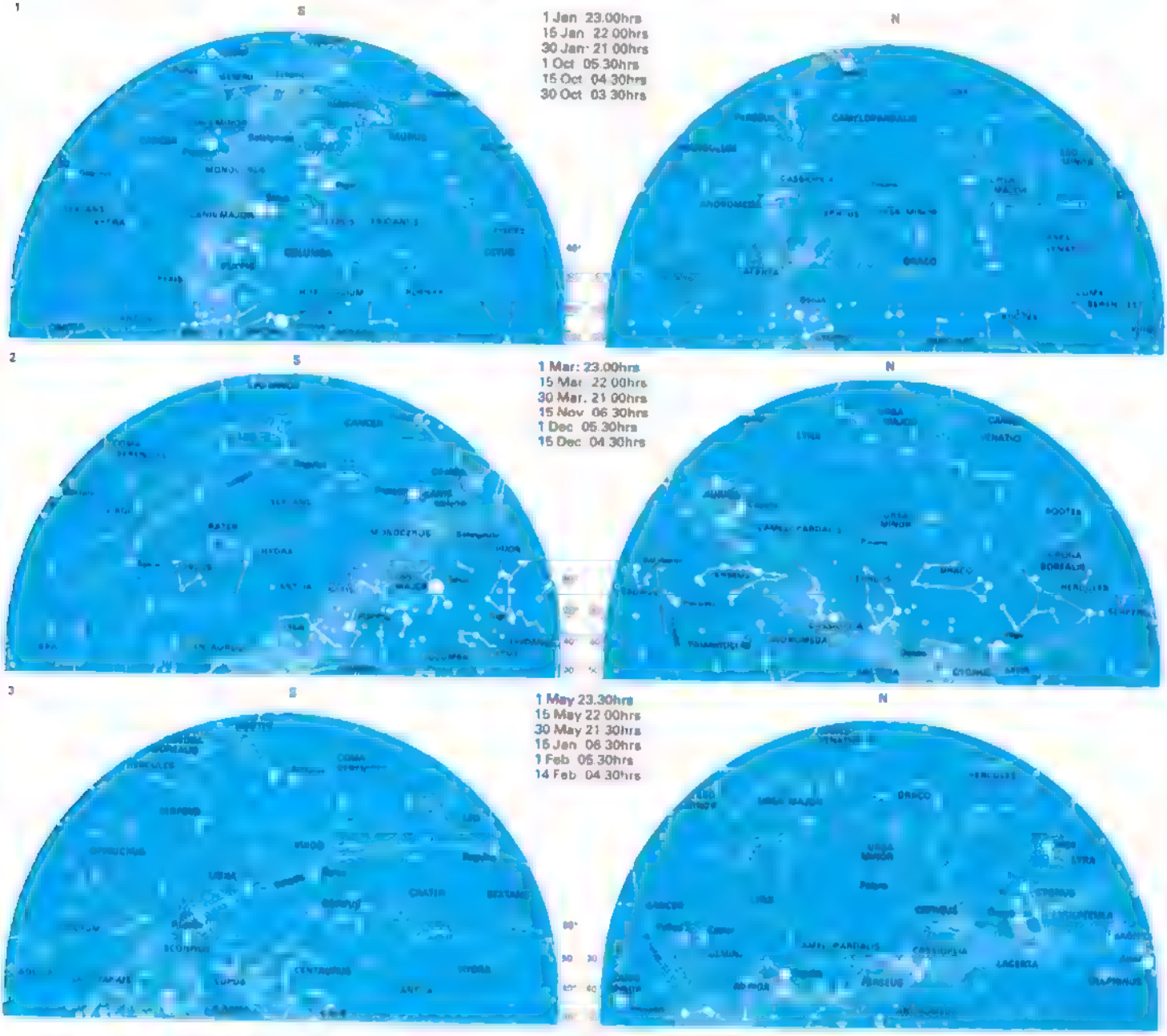
In winter, the southern aspect is dominated by Orion and its retinue. Capella is almost at the zenith, or overhead point, and Sirius is at its clearest – although, as seen from northern Europe or the United States, Sirius is always rather low down and so the effects of the Earth's atmosphere make it twinkle strongly. Other stars in the general region of Orion are Aldebaran, Castor and Pollux and Procyon; while Orion itself has two brilliant leaders, the white Rigel and the orange-red Betelgeuse. Orion is invaluable as a guide to other groups: for instance, Aldebaran can be found by following the upward direction of the three stars in Orion's belt.

The sickle of Leo is prominent in the east

CONNECTIONS

See also

Measuring the nightime sky



(as shown here it is cut by the two maps). In the northeast lies Ursa Major, the Great Bear – the other invaluable aid to recognizing various constellations. The two Pointers show the way to Polaris, the Pole Star (Alpha Ursae Minoris), which is of the second magnitude. It lies less than one degree away from the celestial pole, so that its declination exceeds +89°. Vega is at its lowest and is not shown on the first chart; it is just circumpolar from England, but not from the New York area. Vega, Polaris and Capella lie approximately in a line, with the Pole Star in mid position, thus when Capella is almost at the zenith, Vega is at the horizon (on winter evenings) and when Vega occupies the zenith, Capella can be seen at the horizon (on summer evenings).

Stars of charts 2–6

During spring evenings (chart 2), Orion is still above the horizon, Leo is high up with Virgo to the east; Capella is descending in the northwest and Vega rising in the northeast. In the west, Aldebaran and the Pleiades are still quite clearly visible.

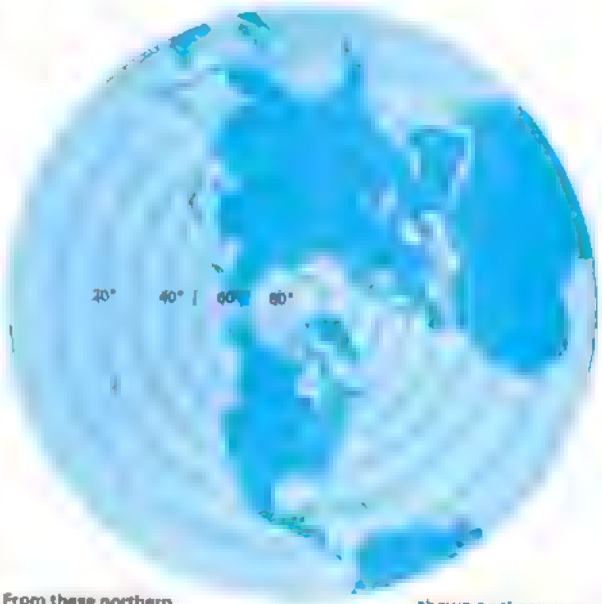
By early summer (chart 3), Orion has disappeared, although Castor and Pollux remain in view. Vega has risen, Capella is descending and Ursa Major is not far from the zenith. This is the best time for evening viewing of Arcturus, which is of a light orange hue and is actually the brightest star in the Northern Hemisphere (fractionally superior to Vega or Capella).

During summer evenings (chart 4), Vega occupies the zenith and its brilliance and distinct bluish colour make it unmistakable. An excellent view is afforded of the so-called Summer Triangle (Vega in Lyra, Deneb in Cygnus and Altair in Aquila) and, in the south, of the brilliant Antares in the Scorpion and the star clouds of Sagittarius in the direction of the centre of our Galaxy.

In autumn (chart 5), the Square of Pegasus is high, the Summer Triangle is still apparent and Ursa Major is at its lowest.

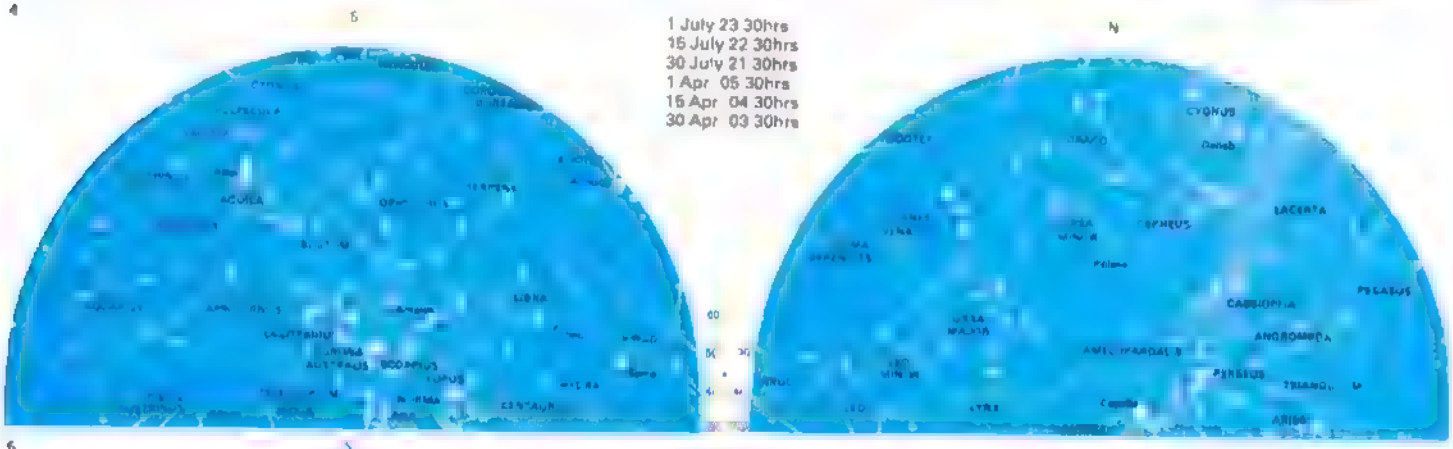
By early winter (chart 6), Pegasus is still high and Vega and its companions are sinking, but Orion has returned once again and will dominate the evening sky until well into the following year.

KEY

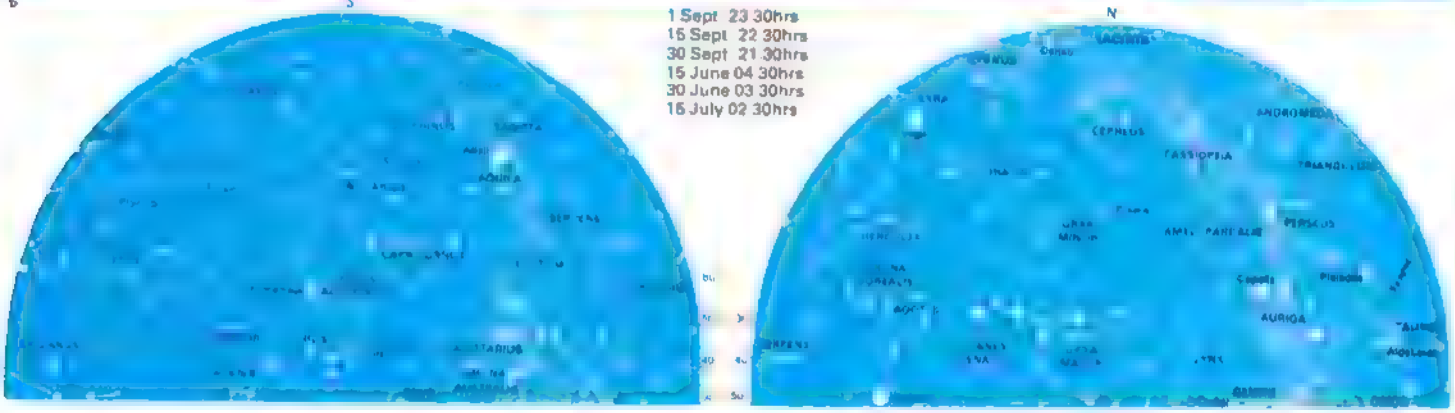


From these northern latitudes the stars shown on these pages can be seen

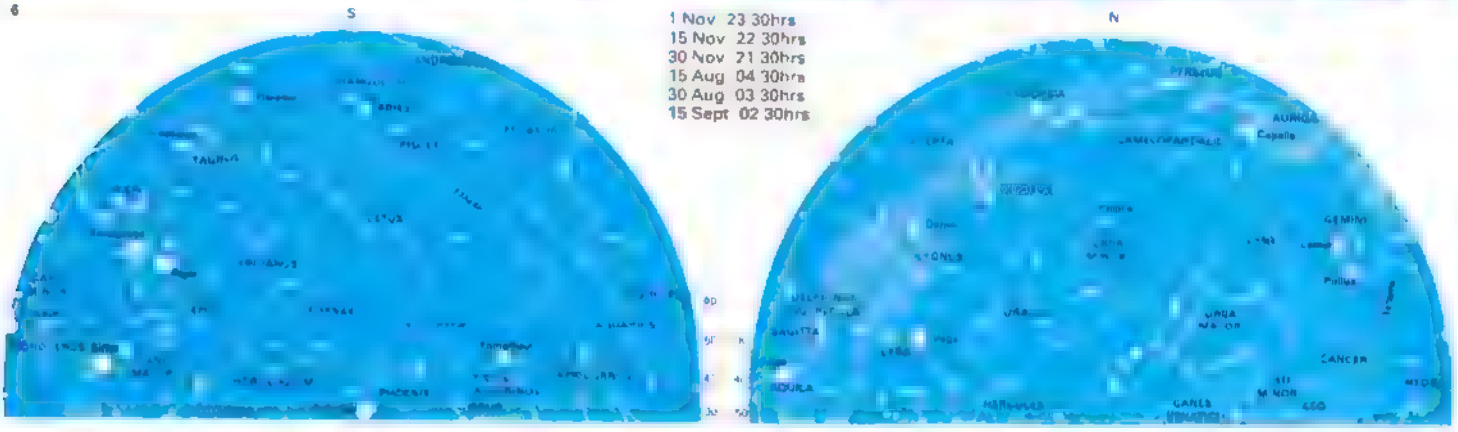
4



5



6



Seasonal star maps: southern

The far southern skies are much superior to those of the Northern Hemisphere for astronomical observation. They contain a number of brilliant groups that cannot be viewed by observers from either most of Europe or the United States.

Northern observers lack both the Southern Cross and the two Clouds of Magellan, which are of special importance astronomically. Other objects of the far south include Alpha Centauri, nearest of the bright stars; the globular clusters Omega Centauri and 47 Tucanae; and the remarkable irregular variable, Eta Carinae. The Milky Way is very rich in its southernmost portions and in the Southern Cross there is the so-called Coal-sack – the best example of a dark nebula – together with the glorious open cluster that is nicknamed the Jewel Box.

Features of chart 1

The six pairs of charts given on these two pages are arranged with the left-hand chart looking towards the northern horizon and the right-hand chart looking towards the southern. They are valid for observers living in South Africa, Australia, New Zealand and most of South America (allowing for differences in aspect) and calculations can be made by reference to a star's declination.

In the evenings during January, the southern summer, Orion is high up, from the Southern Hemisphere Rigel is higher than Betelgeuse, while the Belt stars point "up" towards Sirius and "down" towards Aldebaran. The whole of the Ship (Argo) is displayed; this huge constellation was regarded by astronomers as an unwieldy section for observation and was "cut up", the principal sections now recognized being Carina (the Keel), Vela (the Sails) and Puppis (the Poop). Carina contains Canopus, the second brightest star in the sky, which is very high during January evenings. It does not appear as brilliant as Sirius (its apparent magnitude is −0.7, as against −1.47 for Sirius) but in reality it is much more luminous and lies at a distance of hundreds of light-years from Earth.

Crux is rising in the southeast in chart 1. Strictly speaking it is shaped more like a kite than a cross, it is the smallest of the 88 constellations in the sky, but is very compact. Two of its stars (Acrux or Alpha Crucis and Beta) are ranked as being of the first magnitude. The third (Gamma Crucis) is ranked just below this, while the fourth star of the kite pattern is much more faint. Even a casual glance shows that of the four main stars, Gamma is orange-red (spectrum M) and the others white. The two Pointers to the Cross are Alpha and Beta Centauri; Alpha is the brightest star in the sky apart from Sirius and Canopus and lies at a distance of only 4.3 light-years. It is a fine binary, separable with a small telescope. Alpha Crucis is also a splendid binary.

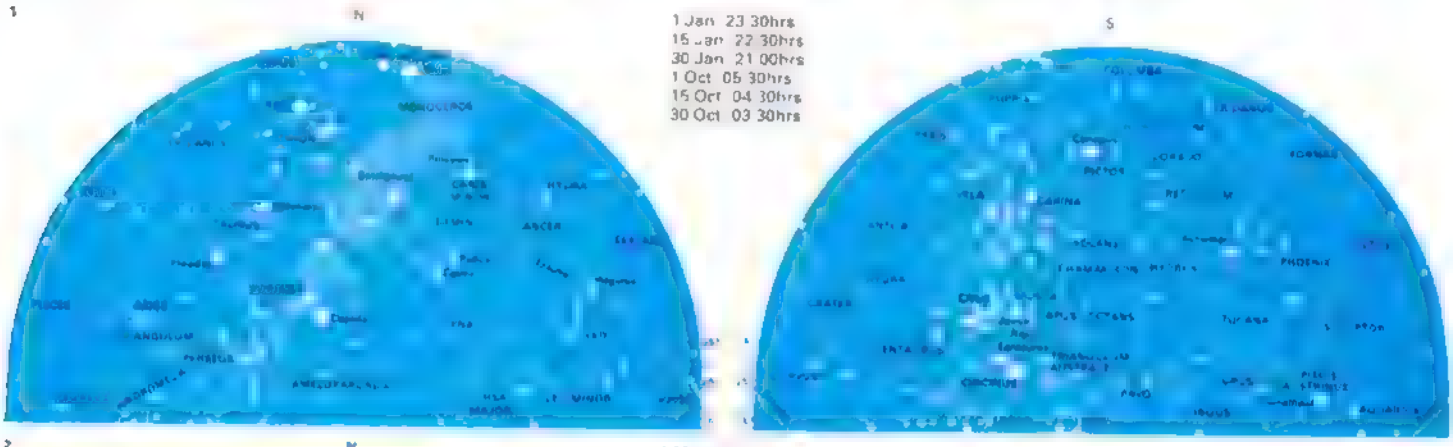
Achernar in Eridanus (the River) is to the southwest. Probably the best way to locate the region of the south celestial pole is to look midway between Achernar and Crux, but there are no really bright stars to help in identification of the pole itself.

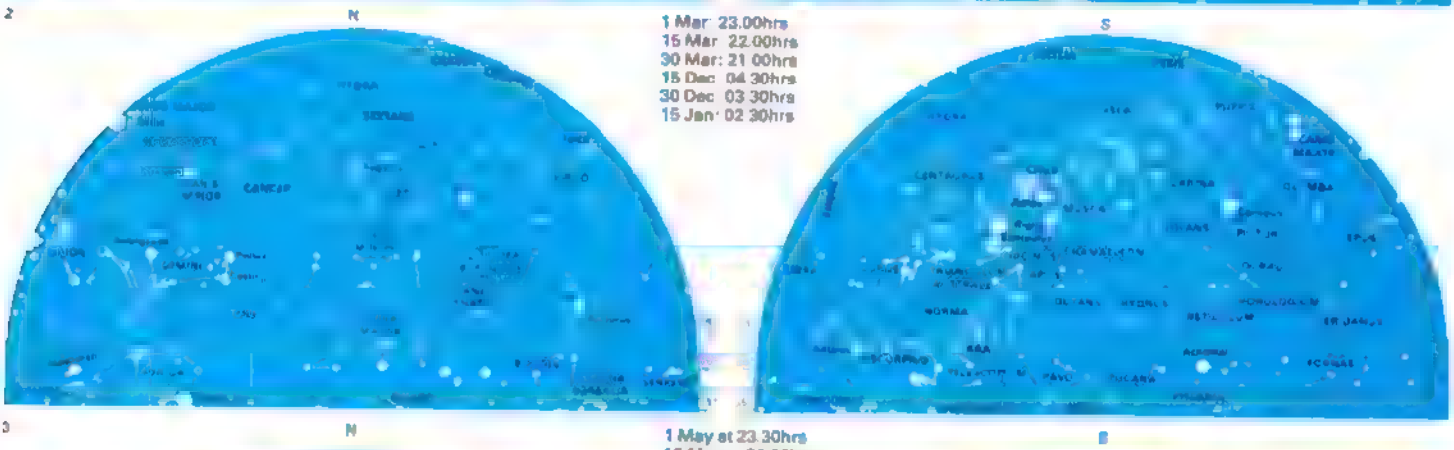
Charts 2 and 3

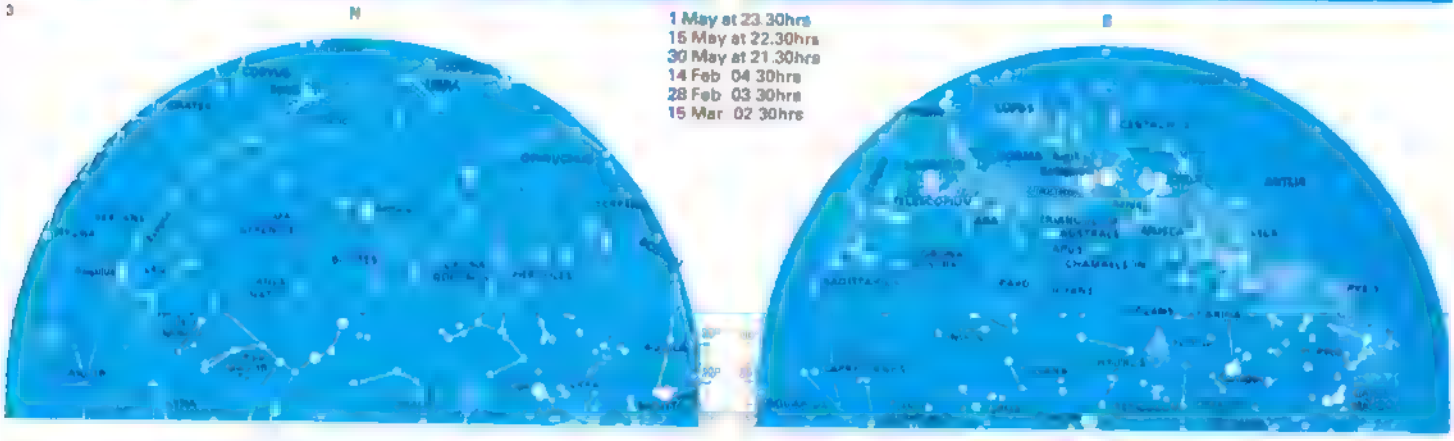
By March evenings [chart 2] Canopus is descending in the southwest and Crux is rising to its greatest altitude, together with

CONNECTIONS

Star guide for northern sky
Star guide northern sky
Star guide southern sky
Star guide southern sky
Measuring the night sky







the Pointers they comprise a splendid group Not far from it is the magnificent globular cluster Omega Centauri, much the finest of its type in the entire sky. The Milky Way is extremely rich in this whole area and even binoculars will give a good view of the Coalsack, which conveys the impression of being a virtually starless area (there are a few stars in the foreground). Scorpio (or Scorpius) is coming into prominence in the southeast, Orion is dropping and Leo can be seen high in the north region

March is the best time of year for seeing Ursa Major, although the constellation never rises to observers in the southernmost parts of the latitudes for which the maps have been drawn. (From New Zealand, for instance, the seven main stars of Ursa Major never rise, but from Rhodesia they can attain a considerable altitude)

May evenings show Alpha and Beta Centauri very high up, with Crux, Canopus is visible in the southwest, but Orion and Sirius have set. Arcturus is prominent in the north, with Spica and Virgo not far from the zenith. The Scorpion is now dominant

The Scorpion group is a splendid one, with a long chain of stars that cannot be mistaken – the red giant Antares is particularly striking. There are many bright, open clusters in the region, as sweeping the sky with binoculars will reveal. Adjoining the Scorpion is Sagittarius, which contains many moderately bright stars.

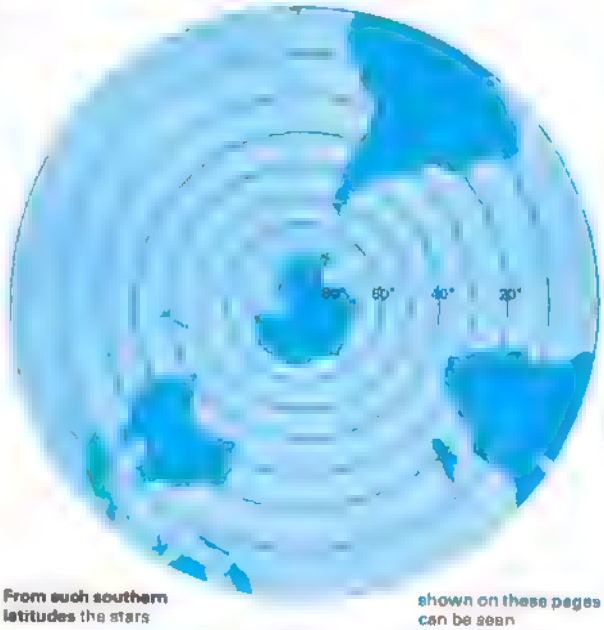
Charts 4, 5 and 6

The Scorpion is near the zenith in chart 4 while Crux and Centaurus remain prominent in the southern aspect of the sky The brilliant northern stars Vega, Altair, Deneb and Arcturus are all visible but Canopus is at its lowest and is virtually absent from the sky for a short while

Pegasus is high in the north in chart 5; Vega, Altair and Deneb can still be seen and Fomalhaut is almost overhead, the Scorpion is sinking in the southwest and Crux almost out of view

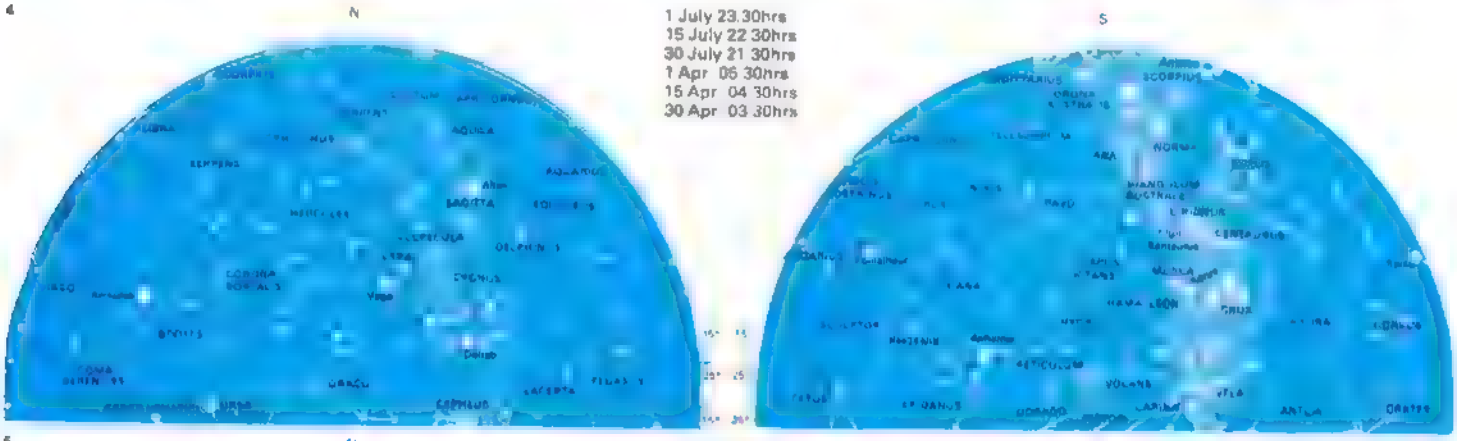
Orion has returned in chart 6 and with it Sirius, Canopus and the other adjacent stars. When Crux and Centaurus also return the southern sky will again be at its brilliant best.

KEY

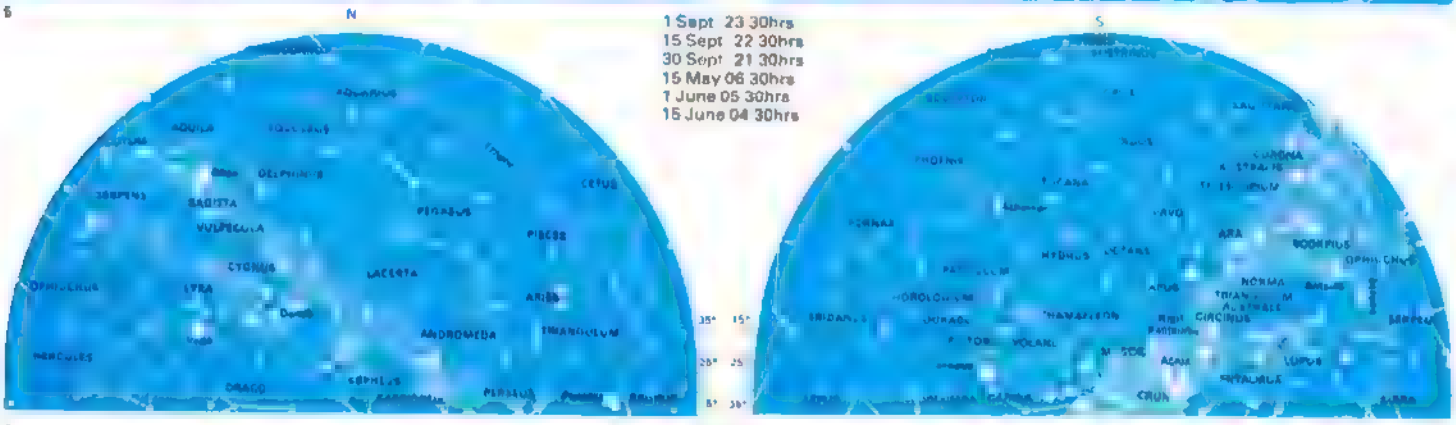


From such southern latitudes the stars shown on these pages can be seen

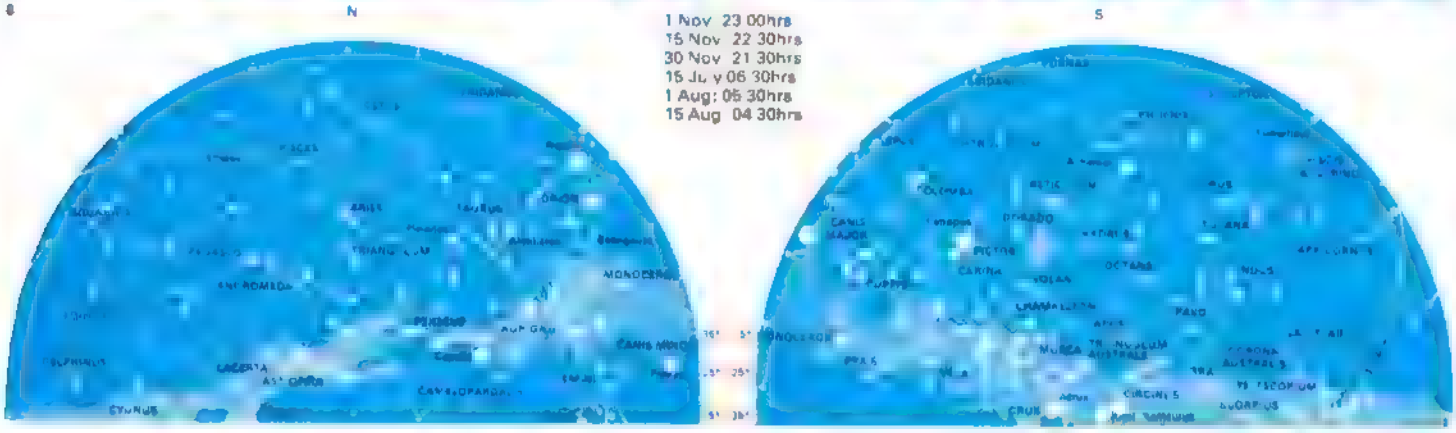
4



5



6



Space vehicles

Since the Soviet Sputnik 1 was launched on 4 October 1957 several nations have between them launched hundreds of artificial satellites into orbit round the Earth The United States and the Soviet Union have sent exploratory "probes" to orbit or soft-land on the Moon, Mars and Venus. Other planets have been studied at close quarters using unmanned space vehicles. The greatest achievement was probably the landing of men on the Moon during the American Apollo programme and there is good reason to suppose that by the end of the twentieth century the whole Solar System will have been explored – but by unmanned probes

Getting into space

The use of vehicles to explore space dates from the development of rockets powerful enough to launch the vehicles and enable them to "escape" from the Earth's gravity [Key]. This requires a velocity of 11 2km per second (approximately 40,000km/h or 25,000mph), which is technically known as the Earth's escape velocity

To reach such speeds, multi-stage rockets are employed These make use of the piggy-back principle [1] – the rocket entering orbit is fired at the edge of the atmosphere, having been carried there on top of another rocket that, in turn, may also have been lifted on an even more powerful first-stage rocket This technique of overcoming gravity was first proposed by the Russian pioneer of rocketry Konstantin Tsiolkovsky (1857–1935). In 1949 an American multi-stage rocket sent a vehicle up to a height of more than 390km (242 miles) above the surface of the Earth

Rockets and satellites

All rockets are reaction motors They work using the principle of Newton's third law of motion which, describing the behaviour of any moving object, states that action and reaction are equal and opposite

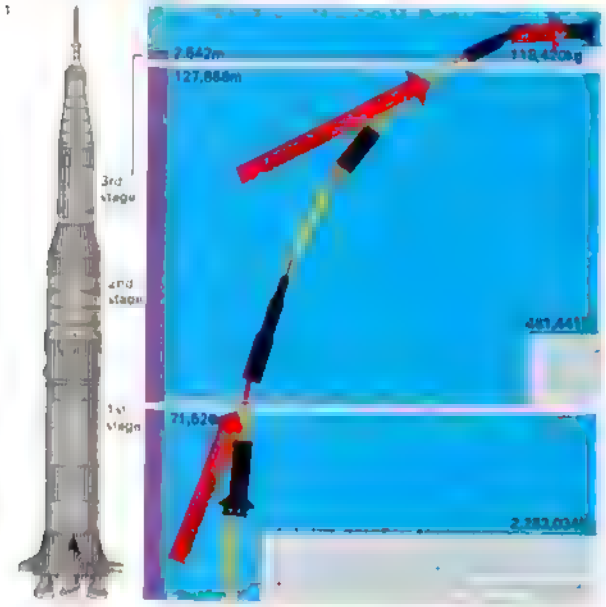
In a rocket the "action" is the escape of hot gases roaring out of the tail, the "reaction" to this action forces the body of the rocket in the other direction. The principle can be demonstrated by blowing up a balloon and releasing it; the action of the jet of air escaping from the neck of the balloon is balanced by an equal and opposite reaction that pushes the balloon through the air For this reason a rocket will work in the airless near vacuum of outer space, it does not, like a jet engine, require a supply of air for the burning of fuel The presence of air is actually a handicap because it sets up a resistance to the rocket's motion

The fuel in a firework rocket is a solid propellant explosive such as gunpowder But solid fuels are too weak and uncontrollable to be used alone in space rockets Instead two liquids are used – a fuel and an oxidant When mixed in a combustion chamber they react together to produce hot gases that are expelled from the exhaust and create thrust The first successful liquid-fuelled rockets were made in the United States in 1926 by Robert Goddard (1882–1945) By the time of Goddard's death German scientists, led by Wernher von Braun (1912–77), had developed the V2. This was a liquid-fuelled rocket that carried a one-tonne explosive warhead and was the direct ancestor of modern space rockets After World War II, von Braun and his colleagues went to the

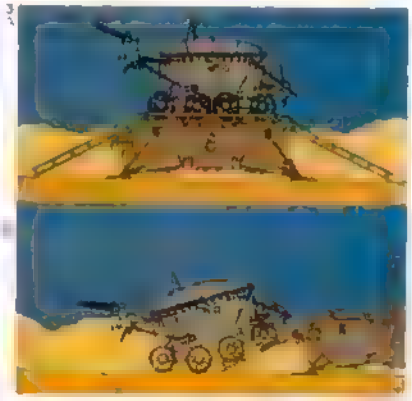
CONNECTIONS

See also

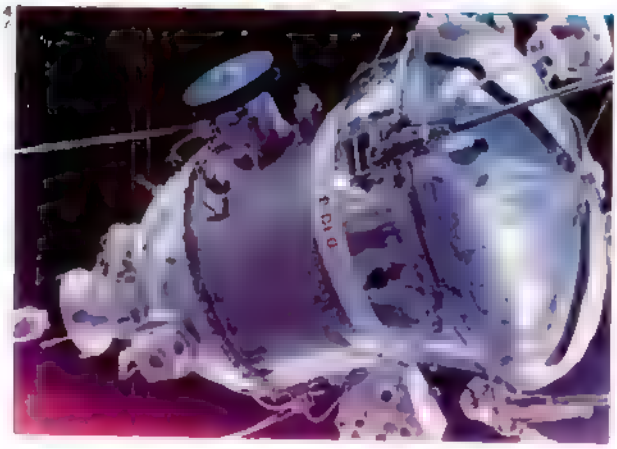
1 Multi-stage rockets consist of a number of smaller rockets combined to make one big one At the start of the flight the large, lower stage is used, here it accounts for 83 3 per cent of the propellant but accelerates the rocket to only 33 per cent of its final velocity When it has used up its fuel, it drops away and the second stage takes over Only the third stage goes into orbit

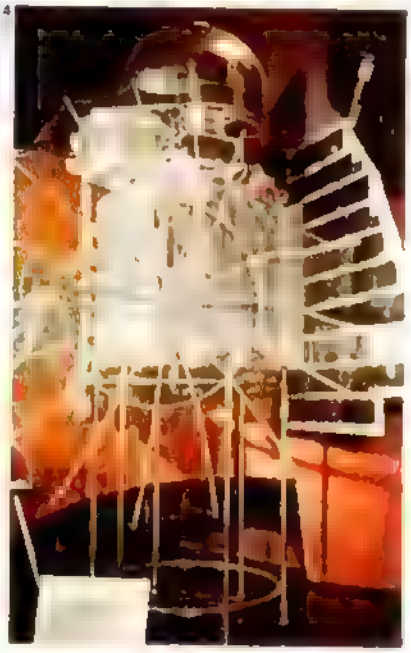


2 The American Vanguard rocket was launched in 1958 in the early days of space research Rockets were not reliable but this craft was a major American success with a minute payload



3 The Soviet "Moon crawler" Lunokhod 1 [A] was taken to the Moon by a Luna rocket probe, after landing it was sent down a ramp [B] on to the surface Lunokhod 1 landed upon the grey plain of the Mare Imbrium and crawled along for months, controlled from the USSR, sending back invaluable data A second Lunokhod operated in the Mare Serenitatis near the landing site of the Apollo 17 module





4 Since 1957 hundreds of space probes have been launched The Soviet Cosmos vehicles [A] are artificial satellites brought back to Earth after limited flight times. Mariner 9 [B] went into Mars orbit in late 1971 Continuing well into 1972 it sent back detailed photographs of the Martian landscape The first probe launched by the Soviet Union to Venus [C] was not a success

United States to continue with their work

Sputnik 1 was the size of a football and carried little apart from a radio transmitter Some of today's satellites are the size of a large truck They have been used in many ways, for mapping [7], communications and scientific research into phenomena impossible to study properly from the ground because of the Earth's atmosphere Communications satellites have evolved from the early passive type, which consisted of a "silvered balloon that acted like a mirror to reflect radio signals beamed up to it back down to Earth, to the modern active satellites that amplify received radio signals before re-transmitting them

Manned satellites have now become relatively common and in 1973 Skylab was placed into orbit as the first true space station Docking procedures have also been carried out between two spacecraft, in 1975 an American vehicle docked with a Soviet one the first such meeting between traditional rivals in space All such manoeuvres require precise information about orbits and velocities. This is provided by radar sets on the craft and on Earth and the necessary complex calculations are carried out using computers suitably programmed for their task

Probes to the Moon and planets

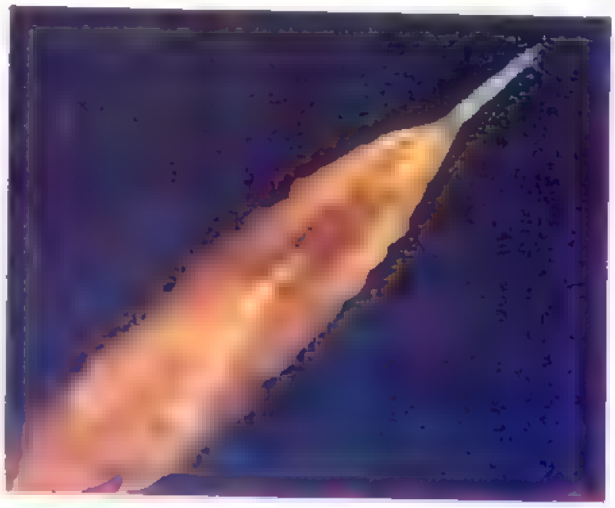
The first target for unmanned space probes was the Moon In 1959 the Soviet Luna 3 made a circumlunar voyage and subsequently the whole of the Moon's surface was mapped by automatic probes. Soft landings on the surface were also made [3] and mechanical fingers used to collect dust and rock

The lunar probes were followed by the first attempts to explore the planets. In 1962 the American Mariner 2 made a fly-by pass of Venus and vehicles have since been sent to Mars, Mercury and Jupiter By 1975 a Soviet probe had soft-landed on Venus and in the same year a vehicle was on its way to Saturn The American Vikings made soft landings on Mars in 1976

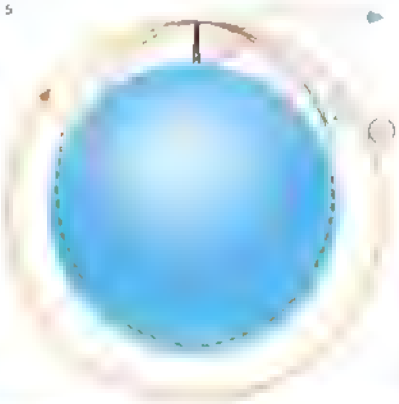
Vast amounts of invaluable information have been collected. Telescopes and spectrometers in space vehicles aid Earth-bound astronomers. And infra-red photographs of Earth can reveal new resources

KEY

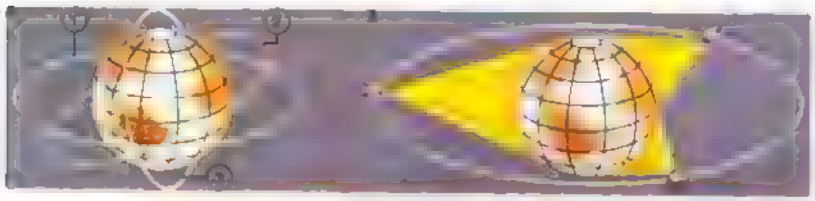
All space vehicles are put into orbit or sent on their journeys to the Moon, other planets and beyond by powerful rockets This rocket launched the American Mars probe Mariner 9 in 1971, carried in the top part of the launcher It was the first probe to be put into a close path round Mars and transmitted back to Earth thousands of high-quality pictures.



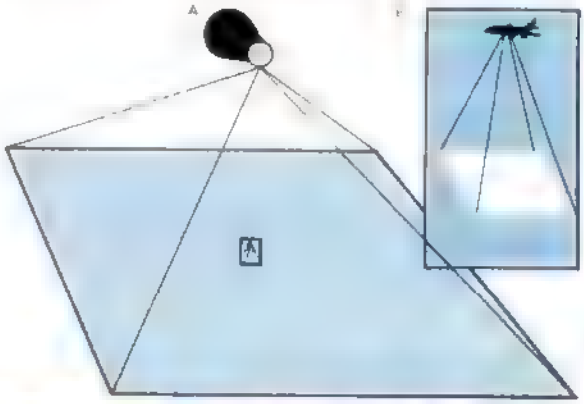
5 Various paths for space probes can be drawn assuming that the vehicles are fired horizontally from the top of a tall tower reaching above the Earth's atmosphere At low velocity [1] the vehicle soon falls back to the ground With greater velocity [2] the vehicle travels farther before landing But with orbital velocity [3] it does not land at all and enters a closed and stable orbit

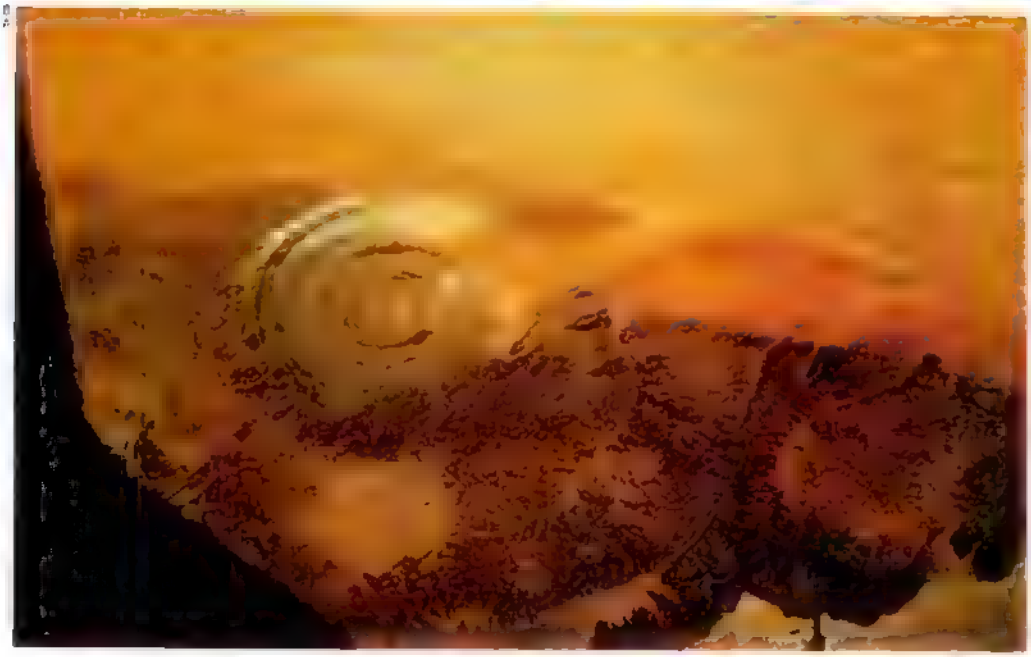


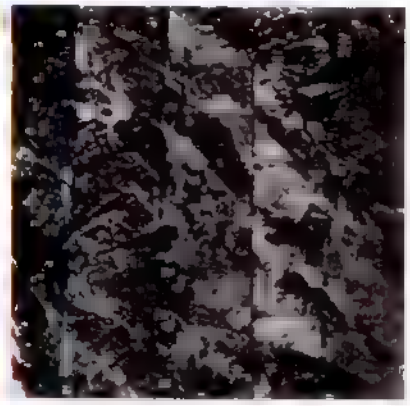
6 Satellites may travel in orbits of various kinds [A, Some move in the plane of the Equator [1], others have inclined orbits [2], and some use polar orbits [3] For a "stationary" satellite of the Syncom communications type [B] the period is exactly one day Its distance from Earth is 35 900km (22 300 miles), it appears stationary and is ideal for television relays



7 An orbiting satellite is better for photographing the Earth than is an aircraft One exposure from a space vehicle [A] can cover an area that would need hundreds of photographs from an aircraft [B] The whole area can be shown with greater accuracy and detail Also, a vertical space photograph does not have the distortion inherent in aerial mosaics of a wide area Aerial photographs require lengthy and specialized processing to make them into a mosaic map whereas this task is greatly reduced using space photographs Any necessary revisions can therefore be made more easily







8 A comparison of space and aerial photography clearly shows the far superior structural definition of the former The Richat craters in Mauritania, West Africa [A], probably of volcanic origin, are well defined in the photograph from the orbiting Apollo 9 vehicle A mosaic of aerial photographs of the same area is shown in [B] The Apollo photograph clearly reveals previously unrecorded features including depressions up to 1 500m (nearly 1 mile) across

Man in space

Yuri Gagarin (1934–68) of the Soviet air force was the first man in space. In April 1961 – less than four years after the launch of the first artificial satellite, Sputnik 1 – he made a complete circuit of the Earth in Vostok 1, above the bulk of the atmosphere before landing safely in a prearranged area.

Gagarin and zero gravity

Gagarin's flight was a truly pioneering venture. Nobody at that time had any real idea of how the human body would react to a prolonged period of weightlessness. Yet during his flight Gagarin experienced conditions of zero gravity – something that cannot be simulated on Earth for more than a brief period.

Zero gravity [1] does not mean that the orbiting astronaut has completely escaped the pull of the Earth's gravity. The best way to picture it is to think of a book placed on a piece of card: the book presses on the card and with reference to it the book is 'heavy'. If both are then dropped the pressure of book on card ceases, the two objects move in the same direction at the same rate.

The same situation occurs when an astronaut is inside his vehicle: the two move at the same rate so that the passenger does not press down upon his craft and 'weight' vanishes. (His mass – the quantity of matter in his body – does not change.)

Gagarin found that zero gravity was neither inconvenient nor unpleasant. This has been confirmed by all later space travellers although "walking" in space [4, 5] is extremely exhausting. The first man to venture outside an orbiting vehicle was the Russian cosmonaut Alexei Leonov, and this has since been repeated many times by both Americans and Russians.

Americans in space

The first American in space was Alan Shepard, who made a sub-orbital flight lasting for about 15 minutes in May 1961. In the 1960s manned satellites carrying two or three astronauts were sent up and there were elaborate docking operations in which two independent spacecraft were skilfully manoeuvred together and joined.

One initial difficulty facing international docking was that American and Soviet designs differed because their space programmes had been developed independently of each other. But following the success of the US Skylab space station plans were made for a joint exercise and this was accomplished with the Apollo-Soyuz mission in 1975. Both vehicles had been suitably modified – the Soviet cosmonauts generally breathe ordinary air at normal pressure, for example whereas the Americans prefer pure oxygen at reduced pressure. For the joint mission a special "adaptation chamber" had to be set up between the two control cabins.

During manned flights many experiments are carried out. The Earth can be closely studied and there have been vast improvements in man's knowledge of the circulation of the atmosphere, which should bring better weather forecasting; plant and mineral resources can be assessed, and nearly all other sciences can benefit.

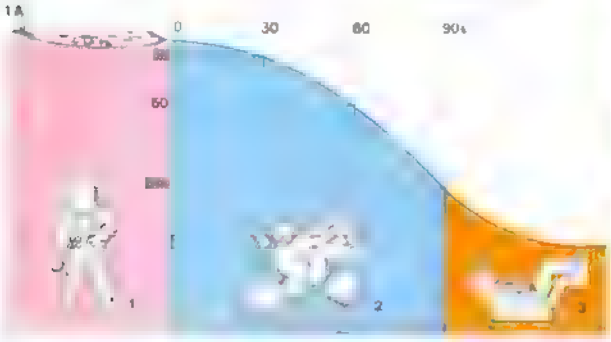
Predicted space-flight dangers from meteoroids (solid particles moving in space), cosmic radiation and weightlessness have failed to materialize. On the other hand mechanical mishap both on Earth and in

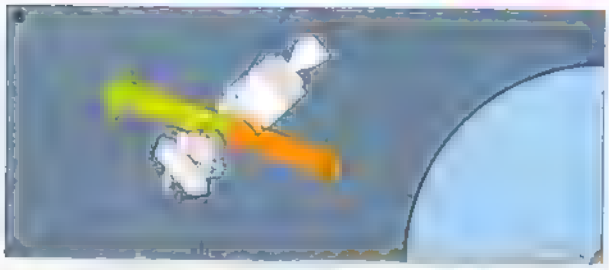
CONNECTIONS

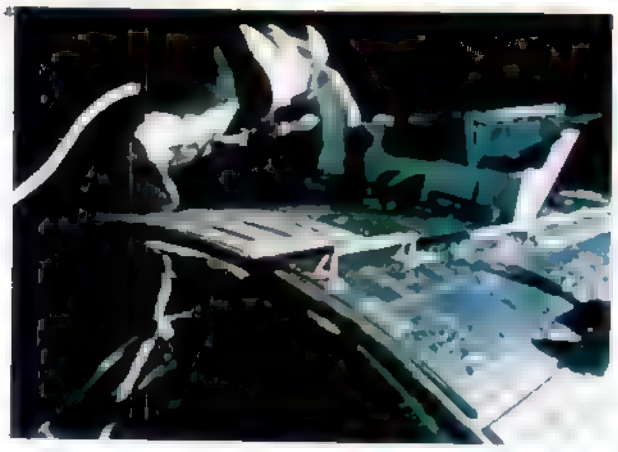
See also

[illegible]

1 Free fall is the condition of weightlessness or zero gravity. [A] shows an astronaut training in a flying aircraft. In [1] he is experiencing normal gravity; in [2] the plane is put into a curving dive that simulates the free fall state for a very brief period; in [3] he takes to a pressure couch to counter the extra g force as the aircraft levels off. [B] shows how in an orbiting capsule, gravitational pull (mg) is balanced by centrifugal force (mv^2/r) to produce zero gravity (m = mass, r = radius of orbit, v = velocity and g = acceleration due to gravity).







2 The Gemini programme followed the first US manned programme (Mercury). Gemini 7, shown here, was able like her sisters to carry two men could conduct docking procedures and allowed for "spacewalking".

3 Apollo 15's command and service modules were photographed by lunar module pilot David Scott. At this time the probe was orbiting the Moon above the Sea of Fertility (Mare Fecunditatis).

4 Astronaut Alfred M. Worden space walked outside Apollo 15 on the return trip he recovered film equipment that had been used earlier. At the Moon he did not go to the surface with Jim Irwin and David Scott but remained in orbit in the Apollo command module.

5 Astronaut Edward White carried out the first US space walk in 1965 during the Gemini programme. (The first spacewalk ever was made by the Russian Leonov earlier in 1965.) Astronauts outside a craft do not drift away but remain in the self same orbit. White was later killed in the fire at Cape Kennedy that destroyed a capsule under test.

space is a danger and there have been deaths of both Americans and Russians. Manned flights to the Moon have so far been the preserve of the Americans. Russians have concentrated on automatic exploration

The Apollo programme [6–9], initiated in the early 1960s, reached its climax with the Apollo vehicles of 1968 and 1969 During the Christmas period of 1968 Frank Borman James Lovell and William Anders orbited the Moon in Apollo 8. In the following year the lunar module was tested close to the Moon's surface Finally in July 1969 Neil Armstrong and Edwin Aldrin landed on the waterless Sea of Tranquillity The gap between Earth and Moon had finally been bridged by man

Inherent problems of Moon missions

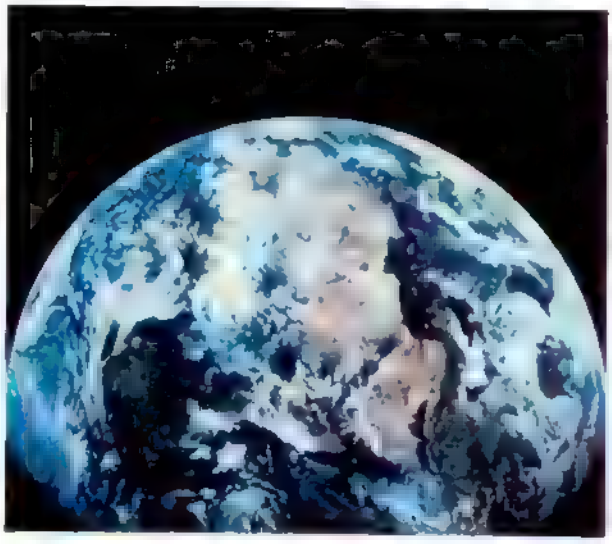
The fuel problem is such that it is not yet possible to send a single-stage vehicle to the Moon and back The initial launching is by step-vehicle, the command and service modules combine then travel to the neighbourhood of the Moon and enter closed orbit Next, two of the astronauts make the final descent in the lunar module, the only function of which is to shuttle its crew from the main spacecraft to the Moon's surface and back Nevertheless the procedure has its dangers. The explorers depend entirely on the ascent engine of their lunar module, if this fails there can be no chance of rescue for the men on the surface

All the landings made so far have fortunately been successful. The only in-flight failure came when an explosion aboard Apollo 13 on the outward journey put the main propulsion unit out of action. The astronauts were forced to use the motors of the lunar module to pass round the Moon and return safely to Earth

Apollos 11 and 12, and 14 to 17 have made great progress in lunar study. ALSEPs (Apollo Lunar Surface Experimental Packages) have been set up and are still operating During the last three journeys the astronauts were able to drive across the surface in lunar rovers or Moon cars Yet the Apollo system was limited in scope: before men can go to the Moon in large numbers there must be provision for rescue, and this may not be possible for another 15 years at the earliest

KEY

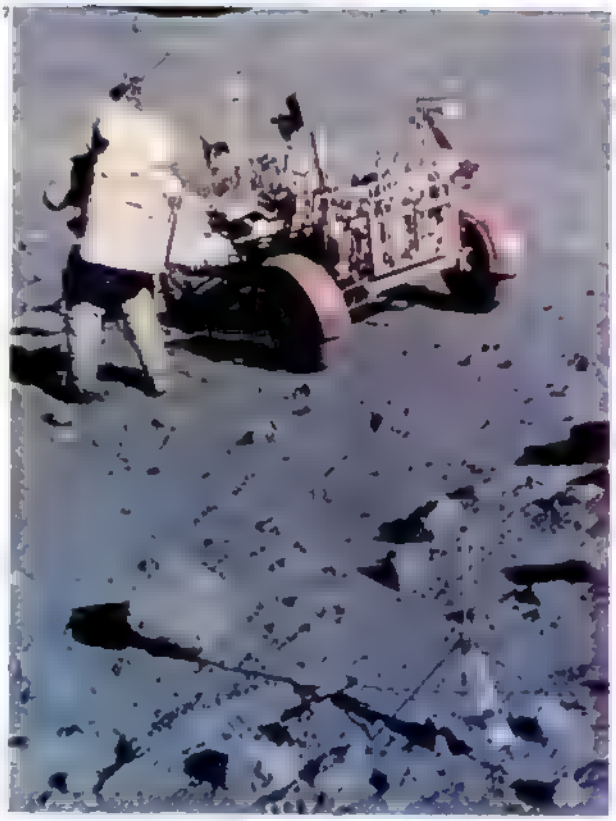
An astronaut's eye view of Earth taken by an Apollo crewman *en route* for the Moon shows both North Africa and Arabia There is considerable cloud cover – but a great part of the Earth was still visible to the astronauts



6 Command and service modules of Apollo 16 orbited the Moon in 1972 Below lay the inhospitable lunar surface with several well defined craters The photograph was taken from the lunar module which carried astronauts Charles Duke and John Young to the Moon's surface

7 Lunar rover vehicles (LRVs) considerably extend the area of exploration for astronauts on the Moon Charles Duke is seen here with the LRV of Apollo 16, near the peak that was soon unofficially named Stone Mountain In the background the bright rays are from South Ray Crater

8 The first Moon landing was made from Apollo 11 in July 1969 Edwin Aldrin stands on the lunar surface filmed by Neil Armstrong who was first to descend the ladder from the lunar module The entire mission – and the 'Moon walk' was shown on TV



9 Hadley Delta, one of the peaks of the lunar Apennines forms the background for David Scott and the Apollo 15 LRV It is farther from Scott than it looks – the distance is more than 30km (19 miles) On the Moon there is no atmospheric scattering so distances can be deceptive and the sky is always black. The US flag does not flutter on its pole since the Moon has no wind – the fabric has to be wired to make it stand out

10 The ascent motor of the lunar module worked perfectly on each occasion that an Apollo craft left the Moon – this is the view from Apollo 15 Yet this was the weakest link in the entire programme If for any reason the ascent engine failed there could be no hope of rescue it is not likely that men will return to the Moon until rescue provision is made The ALSEPs left on the lunar surface are powered by solar cells and are still functioning

The way into space

Although space travel has become a reality only in the closing years of the twentieth century, the means to get into space – the rocket – has existed for more than 1,000 years. Some authorities maintain that the Chinese amused themselves with rocket fireworks as long ago as AD 618, and it is certain that they used rockets as weapons against the invading Mongols during the battle of K'ai-Fung-Fu in AD 1232. Later, the English inventor Sir William Congreve (1772-1828) designed reliable war rockets, fuelled by a highly-explosive mixture of powders, which were used during the Napoleonic wars.

Early rockets

The qualities that make the rocket ideal as a weapon suit it equally for use as a space vehicle. A rocket is propelled by the stream of gas pouring out of its exhaust, and the enormous backthrust on the rocket itself can accelerate it to speeds high enough to break away from the strong pull of Earth's gravity. Because it works on a "reaction" principle (Newton's Third Law of Motion), a rocket does not need air to keep it up or to 'push against'.

All of this was anticipated as long ago as 1895, in the visionary calculations of a shy, partially-deaf Russian schoolmaster Konstantin Tsiolkovsky (1857-1935). He also realized that space rockets would need to be powered by slower-burning, more reliable liquid fuels, rather than solids such as gunpowder, and his proposed combination of liquid hydrogen (as fuel) and liquid oxygen (to make it burn) is used in many rockets today. Robert Goddard (1882-1945) in the US came to very similar conclusions, publishing them as "A method of reaching extreme altitudes" in 1919 – and by "extreme altitudes" he meant the Moon. Goddard followed this up with the first launch of a liquid-fuelled rocket on 16 March 1926. The rocket flew for only seconds before ploughing into his aunt's cabbage patch 55m (184 ft) away.

In Europe, the publication of Herman Oberth's "The rocket into interplanetary space" in 1923 stimulated the formation of Germany's Verein fur Raumschiffahrt, the Society for Space Travel. This bunch of enthusiastic amateurs – among them a brilliant young scientist, Wernher von Braun (1912-) – were soon building experimental rockets which soared to heights of more than a kilometre (0.5 mile). At the same time, the Germany army, under Walter Dornberger, had started a secret project investigating rockets as weapons. Von Braun was invited to join the project in 1932.

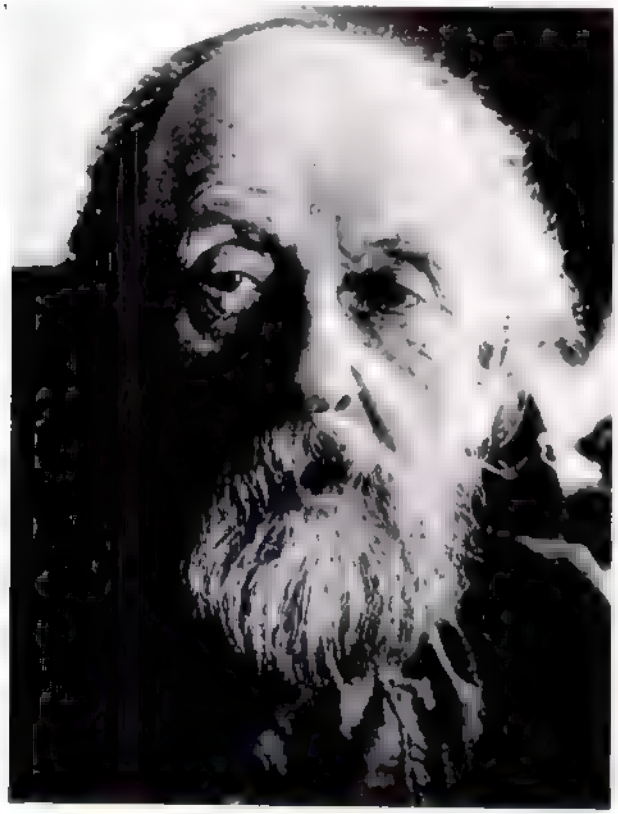
Rocket development

With the outbreak of World War II in 1939, the development of a reliable long-range war rocket became one of Hitler's strongest priorities. The project team launched their successful V-2 weapon, a 4.5-tonne rocket measuring some 14m (46ft) long and 1.5m (5.5ft) in diameter, from the Baltic island of Peenemunde in October 1942. It was fuelled with alcohol and liquid oxygen, and flew on a perfect trajectory to a target 200km (120 miles) away. The first of the 1,500 V-2 rockets aimed towards England hit London in September 1944.

After the war, all 117 Peenemunde

CONNECTIONS

See also



1 Konstantin Tsiolkovsky (1857-1935) photographed in 1930. He was a schoolmaster, shy and partially deaf, but long before the American Wright brothers flew the first plane, he anticipated space flight. His earliest article on the subject published in 1883 very accurately described the weightless conditions of space. In 1895 he put forward proposals for an artificial Earth satellite orbiting at a height of 1[illegible] km (200 miles). In the early years of this century he worked on the fundamental calculations which showed that flight by rocket was possible. Later he went on to demonstrate that the power needed to escape from the Earth would come from building multi-stage rockets, each succeeding stage taking over as the last burned out. Oberth Tsiolkovsky's designs is very similar to a modern Russian Vostok launcher.

2 This Vostok launcher is on display outside Moscow's Cosmos Pavilion. It is the USSR's most widely used launch vehicle and can be varied in its power by the addition of different top stages. The basic craft consists of a central core 28m (92ft) long and 2.95m (9.7ft) in diameter to which are attached four tapered strap-on boosters each 19m (62ft) long. These fall away soon after launch, making the craft lighter and faster.

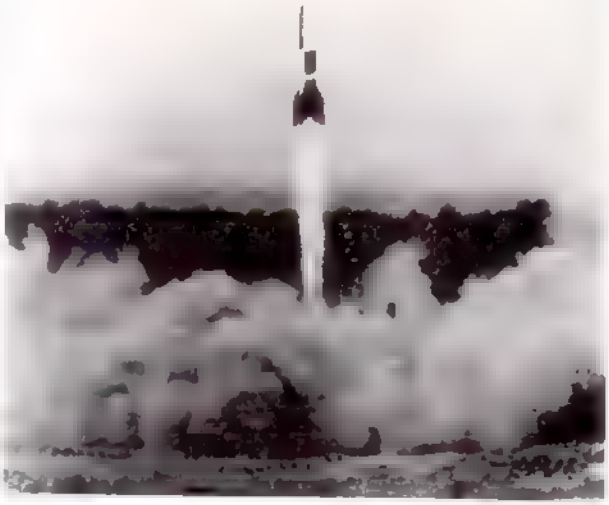
3 Lift-off of an Atlas rocket from Cape Canaveral on 16 May 1963 carrying astronaut Gordon Cooper into Earth orbit for the last of the project Mercury flights. Like the Vostok launcher and many early rockets, the Atlas is a modified inter-continental ballistic missile and it is still used for launches today. The scaffolding at the top of the rocket is an escape tower to pull the astronaut's capsule clear in the event of emergency.

rocket designers and several complete V-2 rockets were taken to the US. The Americans thus had a long lead over the Soviets in commencing their own rocket programme. But the Russians, led by the talented rocket designer Sergei Korolev (1906-66), soon caught up. Both nations concentrated on developing multi-stage rockets – where each successive stage is given a "flying start" by the momentum of its predecessor – with the aim of sending heavy payloads (whether peaceful satellites or nuclear warheads) to great distances. In the event, it was a Soviet multi-stage Intercontinental Ballistic Missile (ICBM), known in the West as an A-type, which launched the world's first spacecraft, Sputnik 1, on 4 October 1957. It was a near thing, because von Braun's US Army team had prepared a rocket and satellite a year before, but were not given permission to launch it.

Space rockets

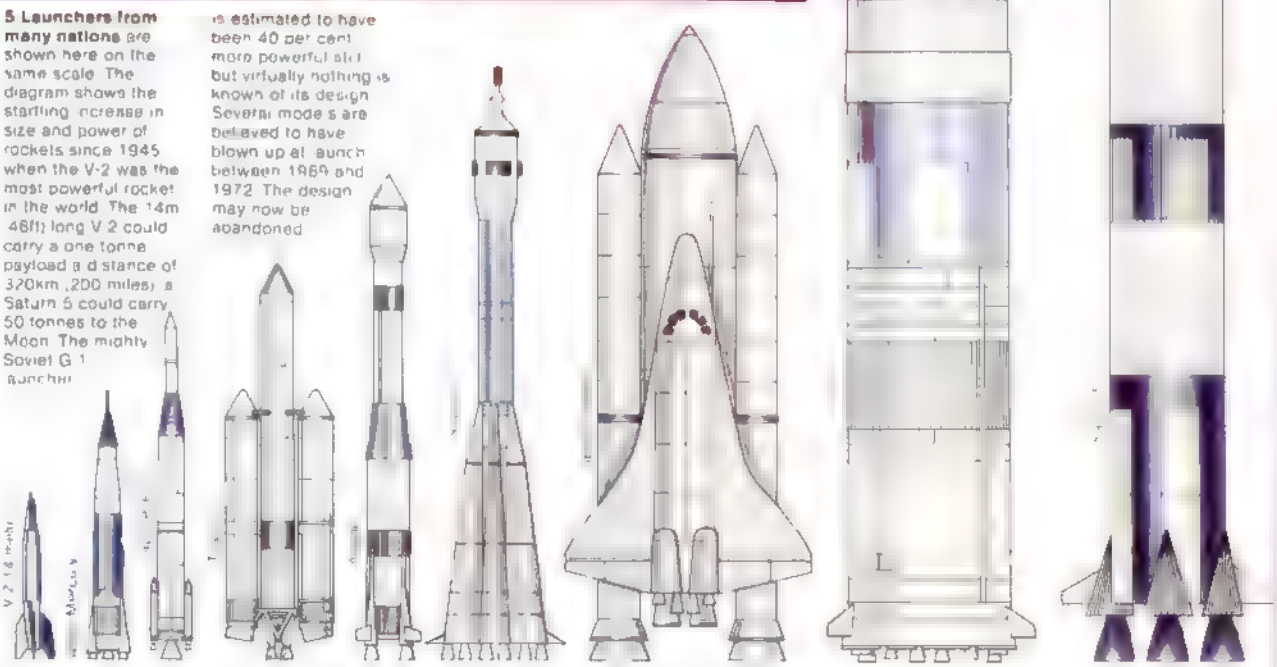
Early space rockets in both the USSR and the US (including von Braun's Redstone rocket) were modified multi-stage ICBMs. As payloads increased in weight, launchers with greater thrust became necessary, culminating in the 111m (363 ft) high Saturn 5 rocket, whose 3,400 tonnes of thrust sent the 43-tonne Apollo capsules on their manned explorations of the Moon. Later Soviet launchers have simply been A-type rockets with additional strapped-on boosters and a more powerful upper stage. the Soviet G-1 superbooster, rumoured to have blown up on its launch-pad in June 1969, is estimated to have been 40 per cent more powerful even than Saturn 5.

Re-usable launchers such as the US Space Shuttle certainly make better economic sense for the future than do expendable rockets. But rockets are still the mainstay of the several countries that feel the need for an independent space programme. China, Japan and India have developed their own launchers, and put their own small satellites into orbit. And the European Space Agency's Ariane rocket, presently capable of launching a 1.7-tonne satellite, should be responsible for the launch of a quarter of all satellites over the decade 1982-91.



This wartime photograph shows the actual launching of a V-2 rocket prototype at a launch-site from the base on the island of Peenemunde.

4 A Saturn 5 rocket lifts three Apollo 8 astronauts towards a journey round the Moon on 21 December 1968. This was the very first manned Saturn 5 launch, all subsequent Moon flights, and the Skylab space station were carried up by Saturn 5 rockets. The 111m (363ft) high launcher, which weighs 2,850 tonnes at lift-off, is able to carry a payload of 150 tonnes into Earth orbit or 50 tonnes to the Moon. Each Saturn 5 rocket was built near the launch site in the colossal 160m (526ft) high Vehicle Assembly Building, a single roomed edifice approximately twice the size of Buckingham Palace. Re-usable launchers are more economical.



5 Launchers from many nations are shown here on the same scale. The diagram shows the startling increase in size and power of rockets since 1945 when the V-2 was the most powerful rocket in the world. The 14m (46ft) long V-2 could carry a one tonne payload a distance of 320km (200 miles); a Saturn 5 could carry 50 tonnes to the Moon. The mighty Soviet G-1 launcher is estimated to have been 40 per cent more powerful still but virtually nothing is known of its design. Several models are believed to have blown up at launch between 1969 and 1972. The design may now be abandoned.

Early years in space

On 4 October 1957 the world's first spacecraft soared into orbit round an astonished Earth. Circling the planet every 96 minutes at a height of between 228 and 947 km (142 and 588 miles) the Soviet Sputnik 1 was a beachball size aluminium sphere that carried no instrumentation other than a 'bleep-bleep' radio transmitter. Western reaction ranged from total incredulity to outright alarm. If the USSR could put a satellite into orbit, the US was no longer safe from the horror of bombardment by Soviet bombs from space.

Up until then, the United State's own satellite programme had been half hearted. A commitment to launching a satellite had been made in 1955, with the aim of putting a scientific satellite aloft during International Geophysical Year (due to commence in July 1957), but there had been little in the way of support on the one hand, or political pressure on the other. Projects of this kind were assigned to military personnel, and the space scientists attached to them were small in number. A hasty attempt to match the Soviet success by launching a satellite with a US Navy Vanguard rocket ended with the rocket blowing up on the launch pad in front of a TV audience of millions. The second Vanguard launch was called off, and in the end it was Wernher von Braun's Jupiter C rocket – developed with a US Army team – which put the tiny Explorer 1 satellite into orbit on 31 January 1958. Although at 14 kg (31 lb) it was only one sixth the weight of Sputnik 1, it was of far more value scientifically. Explorer 1 stayed aloft for six years, and it was responsible for the discovery of powerful zones of radiation surrounding the Earth – the Van Allen belts named for Explorer 1 scientist James Van Allen (1914-).

Meanwhile, the USSR had scored a double first with Sputnik 2, launched on 3 November 1957. This massive satellite weighing in at more than 0.5 tonne (1120 lb) carried the first living cargo – a black-and-white terrier called Laika. Although she was only allowed to live for ten days in space, Laika's survival showed that there were no serious problems facing human space travellers. The main difficulty was in getting them back down to Earth. But the Russians soon demonstrated their capabilities here by recovering the dogs Strelka and Belka alive from Sputnik 5 in August 1960. Not all their flights were as successful, however, the unfortunate dogs on Sputnik 6 both died as their incorrectly-angled spacecraft burnt up on re-entry.

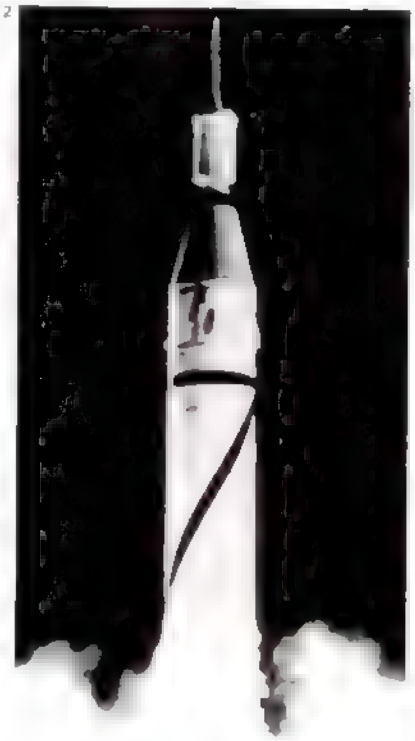
While the USSR's big A type launchers (later to be known as Vostok launchers) were able to put heavy payloads into space, the US concentrated on sheer numbers of launches instead. In 1958 and 1959 they launched 19 satellites, most in the near Earth environment Explorer series. They did not have the capability to put animals into orbit, but instead tested the effects of launch and brief weightlessness on two monkeys, Able and Baker, launched on a sub-orbital "hop" on 28 May 1959. And the US consolidated its commitment to a civilian programme of spaceflight by the creation of the National Aeronautics and Space Administration, usually referred to simply as NASA (and pronounced NASSA), on 1 October 1958.

CONNECTIONS

See also

1 **Laika**, a black and white fox terrier, was the first living creature sent into space. She was put in orbit on 3 November 1957 on board the USSR's half tonne Sputnik 2 and her survival showed that manned spaceflight would one day be possible. Both the US and the USSR carried out further tests with dogs and monkeys before taking a manned mission. Since then a variety of other animals have been into space, including mice, spiders and a wide range of insects.

2

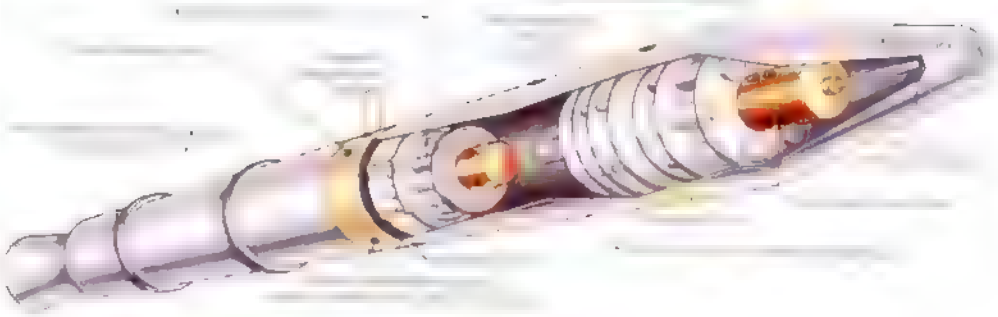


2 The US had their first space success on 31 January 1958 with the launch of the satellite Explorer 1. The Jupiter C launch vehicle, a 17.7m (58ft) high modified IRBM, was the brainchild of a US Army team working under the brilliant Wernher von Braun, who had brought his V-2 expertise to the US after the end of WW2.

3

3 Wernher von Braun stands on the right of this photograph, holding a model of the Explorer 1 satellite following its successful launch in 1958.

4



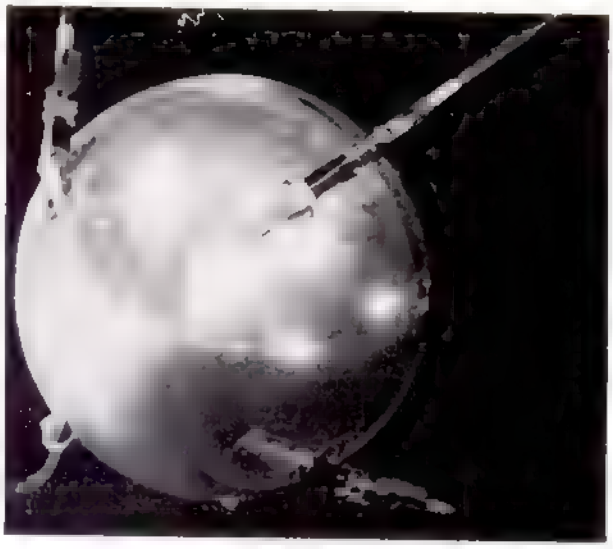
4 This cutaway diagram of the tiny Explorer satellite shows its Sky-Hook scientific instrument, which made the important discovery of the Van Allen belts.

To the Moon

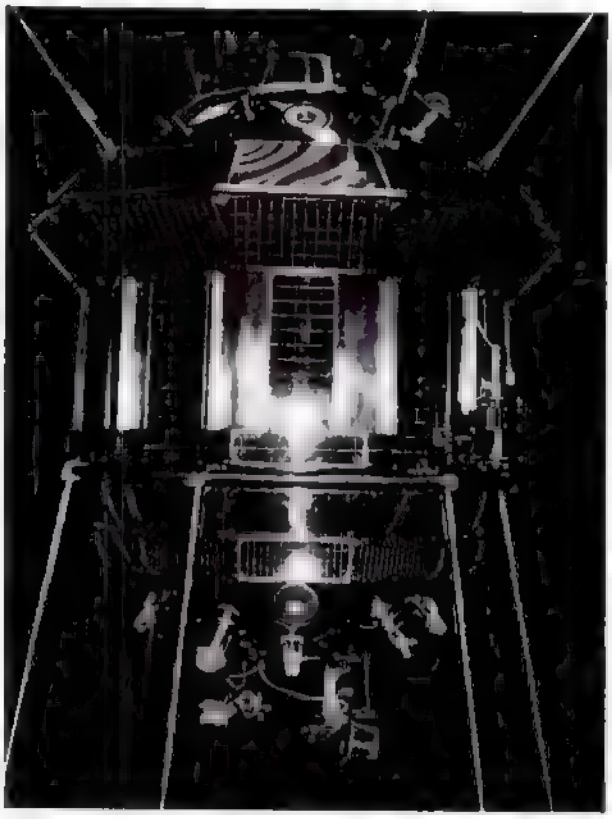
NASA's first space probe, Pioneer 1, was launched on 11 October 1958 towards a very ambitious target – the Moon. The probe, equipped with television cameras, was intended to go into lunar orbit but in the event, it had insufficient thrust and fell back to Earth, mapping the Van Allen radiation belts on the way. The USSR had more luck with Luna 1 on 2 January 1959. Although it missed the Moon by 6,000km (3,700 miles), it became the first "artificial planet", going into its own independent orbit about the Sun, with a year of 450 days. Luna 2 was aimed directly at the Moon and scored a bulls-eye, crashing unceremoniously into the surface near the crater Archimedes on 13 September 1959.

Less than a month later, on 4 October, the second anniversary of Sputnik 1, Russia's third Luna probe grabbed the imagination of the world by disappearing behind the Moon, capturing its permanently-averted far side on film, and returning the pictures to Earth. The technically precise mission and the glimpses of hidden territory – if rather blurred, in retrospect – caused an international sensation. But it was to be Russia's last successful Moon mission for six years, their ten subsequent Moonshots were failures. The US was faring no better with its Pioneer series of intended lunar orbiters: all eight failed, some on the launch-pad, and four of them were not even given series numbers. Only one, Pioneer 5, performed as intended. It was not a Moon probe, but another "artificial planet", orbiting the Sun between the orbits of Venus and the Earth. The Explorer series of Earth satellites was rather more successful, but the US Navy suffered total or partial failure of all but one of their first twelve Discoverer satellites.

Then, amidst all the failure, came the news that took the world by storm. At 6.07 GMT on 12 April 1961, the USSR launched the first man into space. Flight Major Yuri Gagarin (1934-68), aged 27, made one orbit of the globe in his tiny Vostok capsule, and then landed safely near the banks of the Volga. The race into space had begun in earnest and the spirit of competition has persisted ever since.



This 58cm (23in) aluminium sphere is the world's first spacecraft. Sputnik 1 Launched by the USSR on 4 October 1957, it orbited Earth every 96 minutes, transmitting back data on its internal temperatures and pressures.



5 The USSR's Luna 3 was the first space probe to give a glimpse of the far side of the Moon. Because Earth's gravity has captured the Moon's rotation so that it always faces us, we can only ever see half of its surface. Some astronomers in the last century speculated that the Moon's centre of gravity might be displaced towards the far side, thus making it possible for an atmosphere to have developed there. This strange theory was never taken seriously but it was certainly true that there was no positive information about the [illegible] side of the Moon [illegible] visible from Earth in 1959. These [illegible] were [illegible] completely displaced [illegible] on 4 October 1959. Luna 3 was the first Moon probe to carry [illegible]. For a period of 40 minutes it photographed the Moon's hidden face revealing an even more battered and [illegible] surface than the [illegible] and just as desolately forbidding.

6 The two Telstar satellites, launched by the US in July 1962 and May 1963, were the first attempt to use satellites for communications. Telstar 1 [illegible] the first live transatlantic television broadcast as well as dealing with telephone and telegraph transmissions. Although they were an immediate commercial success the problem lay in tracking them: they orbited Earth roughly once every three hours and thus had to be continuously followed by ground station antennas. Syncom 2, shown here, was a great advance on the Telstars because it did not have to be tracked. It was launched by the US on 26 July 1963 into a geostationary orbit 42,000km (26,000 miles) high at which its orbital period matched the Earth's spin and the satellite stayed fixed, as it were, in the sky.

7 Only 384,400km (238,000 miles) distant, our Moon was the natural target for the space race of the 1960s. However, a great deal of unmanned reconnaissance needed to be done before any rocket could carry the first manned expedition. This photograph shows the Moon as it appears when seen through a small telescope. The low lying dark *maria* were thought to be seas by the early telescopic observers, but we now know that the huge impact scars flooded with lava from the Moon's interior. The paler upland areas are covered in craters of all sizes, smaller craters of meteorite hits from space. Before probes went to the Moon, a few scientists worried whether its surface might be covered with a fine dust so deep that spacecraft would sink in it. Others wondered if it might still be geologically active. But now men have been there and brought back rock samples, we know without doubt that it is utterly dead.

The Space Race

"I believe that this nation should commit itself to achieving the goal, before the decade is out, of landing a man on the Moon and returning him safely to Earth." With these words, spoken before the US Congress on 25 May 1961, President John F Kennedy fired the official starting pistol to begin the "Space Race" It was an absurdly daring commitment America s first astronaut, Alan Shepard (1923-), had made his maiden flight into space only 20 days before Instead of orbiting Earth, as the Soviet cosmonaut Yuri Gagarin had done in his flight a month earlier, Shepard's journey was just a quick space-hop

The conical Mercury capsule was however designed for Earth orbit. After one further sub-orbital test, John Glenn (1921), cramped in his telephone box sized cabin, completed three fraught orbits on 20 February 1962. The whole mission was marred by an equipment malfunction which threatened that the craft's heat shield would disintegrate on re-entry But Glenn splashed down safely in the Atlantic Ocean

All this was small currency to the USSR. They followed up Gagarin's first mission with a spectacular series of Vostok flights In the years 1961 and 1962, six Russian cosmonauts – one a woman, Valentina Tereshkova – logged a total of 259 Earth orbits. Over the same period, six Mercury astronauts managed only 34 Although the smaller Mercury craft was more sophisticated than its Soviet counterpart, allowing the astronauts much greater control, America's track record in space looked very poor

The situation reversed dramatically between March 1965 and November 1966 In their roomier two-man craft of the new Gemini series, the US astronauts travelled a grand total of 28 million kilometres (17 5 million miles) in space, compared to the cosmonauts' 12 million kilometres (7 5 million miles). They practised close approaches between space vehicles, and later, docking – essential for the lunar missions By long-duration flights and extravehicular "spacewalks" (first pioneered by the Soviet cosmonaut Alexei Leonov [1934-], they tested the effects of weightless conditions on the human body: no contraindications were found Meanwhile, the unmanned US Ranger and Orbiter craft were scrutinizing the Moon's surface, looking for suitable landing sites. By November 1966, the US was preparing to journey to the Moon

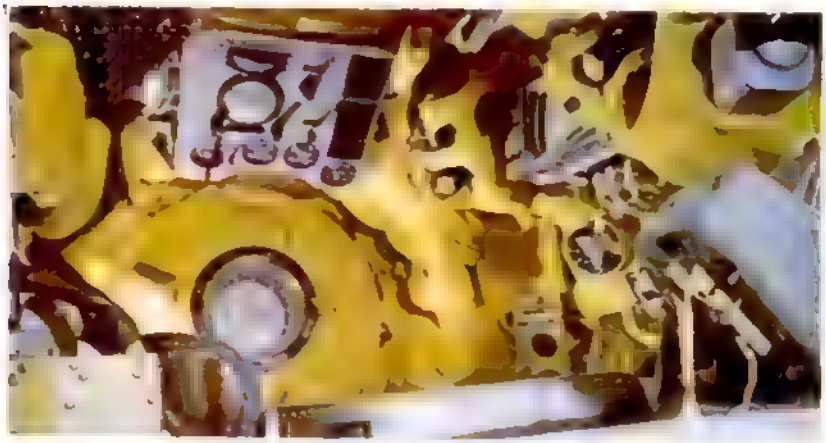
Tragedy strikes

Already, three of the Apollo Moon-craft had been tested unmanned in near-Earth orbit. The project now awaited manned tests and the completion of the colossal Saturn 5 launcher needed to take three men and their lander to the Moon But on 27 January 1967 a sudden fire broke out in an Apollo capsule during a simulated countdown. Astronauts Virgil Grissom, Edward White and Roger Chaffee stood no chance of survival. Not only was the disaster a tragedy on both human and national levels, it also highlighted serious design faults in the Apollo capsule. It was nearly two years before any more Americans flew in space

The Russians, too, were suffering severe setbacks. Sergei Korolev (1906-66), head of the Soviet space programme, had died suddenly in 1966 at the age of 59. His plans

CONNECTIONS

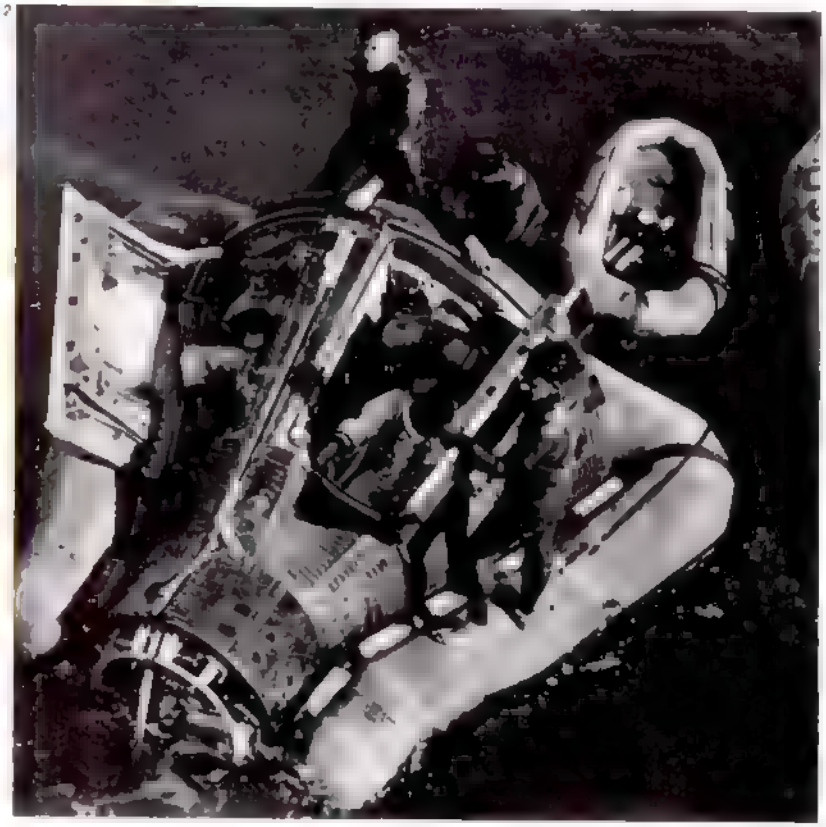
See also



1 Vostok, whose interior is shown here was the first Soviet manned space craft Between April 1961 and June 1963 craft of this type launched by Vostok launchers carried six cosmonauts into Earth orbit The craft s life support systems could maintain a single cosmonaut for up to ten days In the event the longest flight was of five days duration made by Valery Bykovsky in June 1963 when he covered 81 orbits of the Earth Vostok was basically spherical in shape with a weight of about 2 400kg 5 300 lb) and a diameter of 2 3m (7 5ft) Instead of the pure oxygen atmospheres of the early US spacecraft its cabin contained the far safer mix of ordinary air at normal pressure But despite its greater size and weight than the corresponding US Mercury capsule Vostok had few onboard controls and was little more than a passenger craft The cosmonaut sat in an ejector seat with a porthole in front of him so that he could escape in the event of a launch emergency Once in space there was little the cosmonaut could do to control the craft it had no steering rockets and relied on a balance arrangement to align it automatically for re entry into the atmosphere Vostok craft were put into deliberately low orbits so that atmospheric drag would bring them down after ten days The American counterpart allowed more control

2 While the USSR have always preferred their cosmonauts to make soft landings (with the aid of parachutes) over Soviet territory, it has been standard US practice for astronauts and their craft to splash down in the ocean This photograph shows a US Navy recovery of a two-man Gemini capsule A dozen of these craft orbited Earth between 1964 and 1966 giving astronauts the opportunity to practise manoeuvres in readiness for Apollo missions

3 The USSR's Soyuz craft has for long been its standard orbital vehicle and under the name of Zond it was used for unmanned test flights to the Moon during the "Space Race" At the front is a nearly spherical compartment measuring 2 6m x 2 25m (8 7 ft x 7 3ft) where the cosmonauts work in space They sit in the central bell-shaped command module during take-off and re entry while the service module houses supplies and manoeuvring rockets



for the Moon hinged around a new manned spacecraft called Soyuz, which had not yet been tested. Korolev's less-experienced successors were apparently not up to the task of carrying through the first test flight. Only three months after the US space tragedy, Soyuz 1 plunged prematurely to Earth out of control, killing veteran cosmonaut Vladimir Komarov (1927 67)

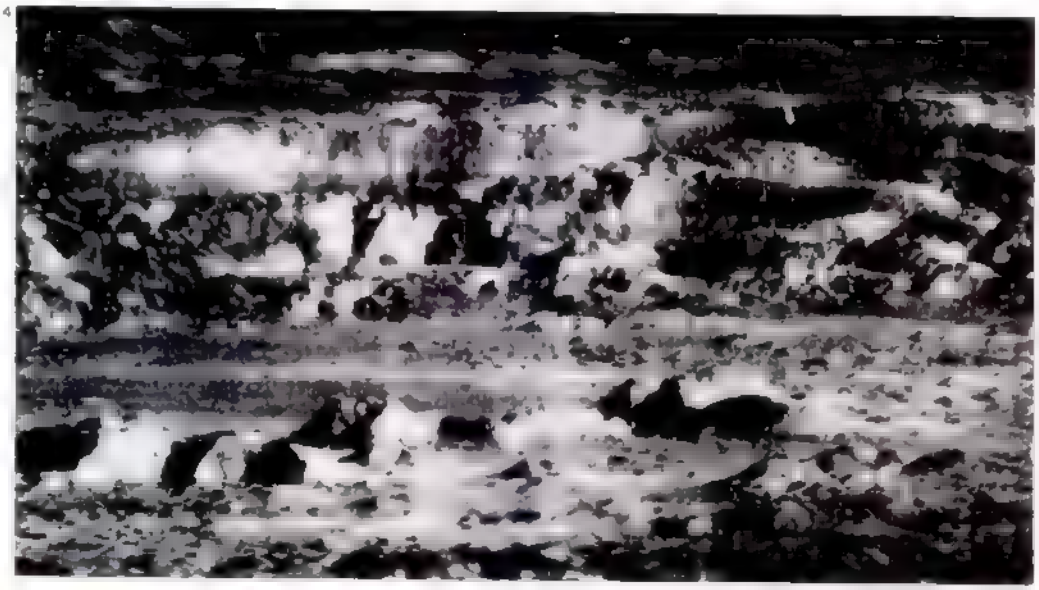
The final lap

While the USSR recommenced its Soyuz programme in October 1968 with an unsuccessful attempt at docking Soyuz 2 and 3, the US comeback was nothing short of spectacular. Following a manned test of the Apollo 7 capsule in Earth orbit, Christmas 1968 saw astronauts Frank Borman, James Lovell and William Anders circling the Moon at a height of only 111km (69 miles) in Apollo 8 – the first manned craft to be launched by the mighty Saturn 5. With this daring mission, the US convinced the world that the Moon was within its grasp

Two further Apollo rehearsal flights took place Apollo 9 (March 1969) in Earth orbit and Apollo 10 (May 1969) round the Moon, the latter a full dress rehearsal for the lunar landing, with the landing craft descending to within 14km (9 miles) of the lunar surface. The stage seemed set But the USSR was not yet beaten Although the deaths of Korolev and Komarov and the explosion of their ultra-powerful G-1 launcher on its launch-pad had put paid to their hopes of a manned expedition to the Moon, they could still be the first to bring back Moon-rocks – and claim the propaganda victory that they had not risked human lives With three days to go before the launch of Moon-bound Apollo 11, the USSR sent the unmanned Luna 15, with a drilling rig and return craft, to the same target

It was an audacious try, but it failed Luna's retro-rockets cut out too soon and the craft plummeted to the Moon's surface on 21 July 1969, exactly the same day that Apollo 11 commander Neil Armstrong stepped down from the Eagle lander to become the first man on the Moon. Kennedy's pledge was fulfilled, and the US had, for the time, won the Space Race

Yri Gagarin (1934 68) was the first man to fly in space His single orbit flight on 12 April 1961 was to be his only trip into space for he was killed in a plane crash on 27 March 1968 while training for another mission. A year [illegible] Soviet [illegible] that Komarov died [illegible] Soyuz 2 mission



4 Hailed as "the picture of the century" when it first appeared in 1966 this oblique view of the lunar crater Copernicus was for most people their first close-up look at the Moon It was taken by the US unmanned space probe Orbiter 2, one of five probes sent into Moon orbit between 1966 and 1967 to photograph the surface in detail and look for possible landing sites The region around Copernicus is too rough for a manned landing. It is scattered with rubble from the impact that created the crater a mountain-ringed hollow 93km (58 miles) across

5 Less than three years after the Copernicus photo was taken astronaut "Buzz" Aldrin faced cameraman Neil Armstrong for this historic shot from the first manned mission to the Moon Armstrong, and the Eagle lander from the Apollo 11 mission can be seen reflected in Aldrin's visor Aldrin's backpack which carried enough oxygen for 4 hours weighed 49kg (108lb) alone the total cost of his Moon suit was in the region of $100 000 Both men spent 2½ hours on the Moon's surface and brought back to Earth some 22kg (48 lb) of lunar rock samples

6 The Soviet alternative to manned Moon missions was to send an automatic lunar car Lunokhod, to scour the surface instead Lunokhod 1 was delivered by the Luna 17 soft-lander in November 1970 The 2 2m (7 25 ft) long "car" rolled out of the lander on eight wheels and began to transmit television pictures of its surroundings Other instruments on board detected cosmic rays, analysed the physical and chemical composition of the lunar rocks, and bounced laser beams back to Earth so that scientists could determine the precise distance to the Moon Lunokhod 1 covered a distance of 10 5km (6 5 miles) on the Moon's surface Lunokhod 2 covered an impressive 37km (23 miles)

Using space

Manned spaceflights and planetary probes such as the Voyagers are only the glamorous tips of a huge iceberg. Most of the rocket launches are intended to put small unmanned satellites into orbit round the Earth, to use near-Earth space for the benefit of mankind.

Communication satellites

A communications satellite (Comsat) can receive a signal from anywhere on one half of the Earth and re-transmit it to anywhere else in the same hemisphere. With a comsats network, signals can be transmitted world-wide.

The US experimented with a Comsat called SCORE in December 1958, less than a year after the first successful satellite launch. They also tried *passive* communications satellites, reflecting radio waves off two huge metallised balloon-like satellites called Echo 1 and 2 (launched in 1960 and 1964). But these were soon abandoned in favour of *active* satellites, with electronic circuits powered by solar cells to receive, amplify and re-transmit the signals. The first commercial Comsat, Teistar 1, was launched in July 1962, and was an instant success in transmitting messages between Europe and the United States.

The USSR launched the first of its series of Molniya comsats in April 1965. These follow very elongated orbits, inclined at 65° to the equator, so they spend three-quarters of their orbital period above the USSR.

With these satellites, the transmitting and receiving aerials must move to follow the orbital motion. Back in 1945, however, British science writer Arthur C. Clark had pointed out that a satellite in an orbit, 35,900km (22,300 miles) above the ground, will orbit the Earth once in 24 hours. Since the Earth itself spins round once in this period, such a satellite will always appear to hang over a particular point on the Earth's equator. A fixed receiving or transmitting aerial can then be used, with a much lower cost and greater reliability.

The first commercial satellite to use Clarke's *geostationary* orbit was Early Bird, launched in April 1965 and positioned permanently above the Atlantic. Geostationary orbits are so convenient that virtually all modern communications satellites use them, including the latest Soviet Statsionar series, successors to Molniya. By the end of the 1980s, there should be more than 150 geostationary Comsats in operation for both international and domestic communications.

Some of the new Comsats will link not telephone exchange or television centres, but individual private users. The Canadian government already operate a series of Anik satellites which communicate with isolated settlements in the Northern Territories.

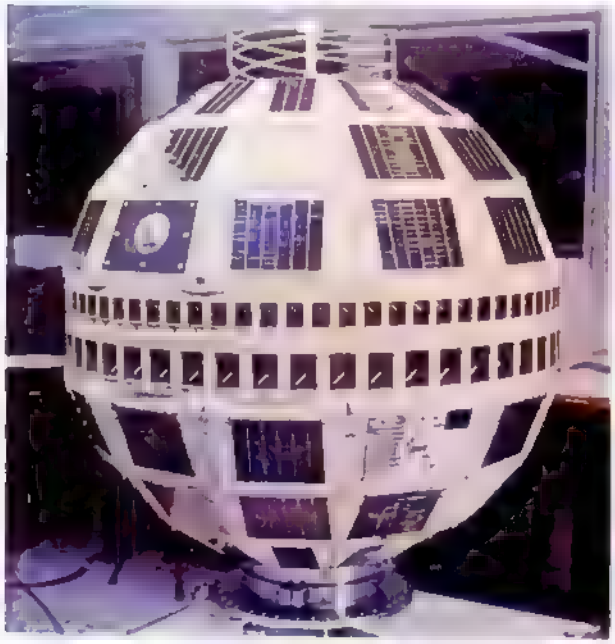
Watching the Earth

From their elevated positions, satellites are ideally located to view the Earth, to monitor weather and analyse the Earth's resources.

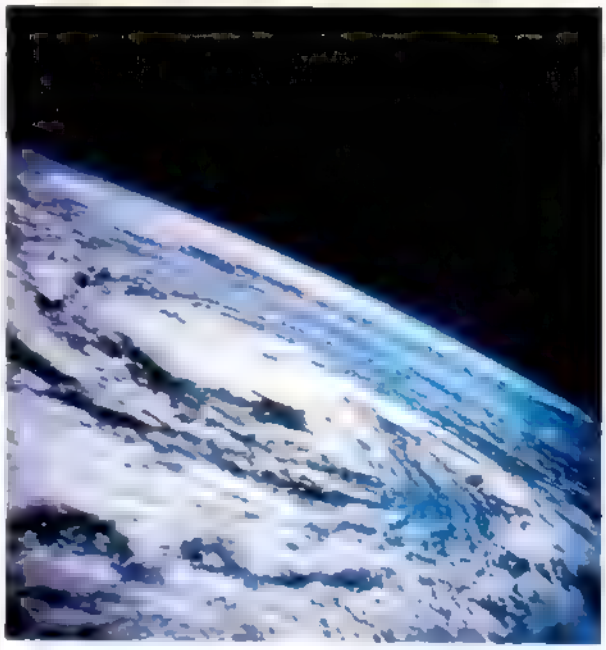
The first weather satellite, Tiros 1, was launched in April 1960. Among its successors are the NOAA satellites (launched 1970-74), which have polar orbits only 1,500km (900 miles) high to give them maximum coverage of the Earth at very high resolution. The US GEOS meteorological satellites (1975) and the European Meteosat (1977) are in the more-distant geostationary

CONNECTIONS

See also



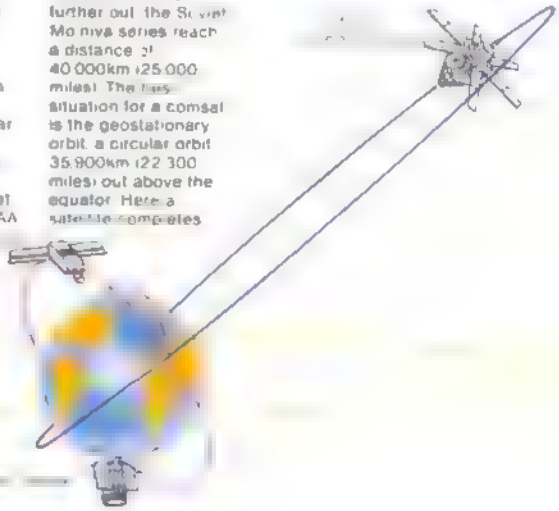
1 Telstar 1, the first transatlantic communications satellite, was 81cm (32 in) in diameter and weighed only 77 kg (170 lb) - less than one tenth the weight of modern Intelsat V comsats. Launched on 10 July 1962, to an orbit that carried it round the Earth once in 158 minutes. Telstar relayed its first broadcast on 23 July.



2 Hurricane Gladys, as seen by the Apollo 7 astronauts in October 1968. Unmanned meteorological satellites now routinely photograph the Earth's cloud cover. In addition to weather forecasting, they have saved hundreds of lives by predicting the likely courses of hurricanes.

3 The European Space Agency's OTS-2 communications satellite undergoes weighing and balancing tests before its launch in May 1978. OTS (Orbital Test Satellite) weighs nearly half a tonne and can handle eight TV channels or 7 200 telephone calls. It is the prototype for an improved version.

4 Satellites are put in different orbits according to their function. Earth resources satellites, like the US Landsat series, have low orbits (900km 550 miles), to achieve maximum resolution of detail. The Landsats are in polar orbits so they can scan all parts of the Earth from directly overhead. The Soviet Meteor and US NOAA weather satellites follow similar polar orbits. Communications satellites can be further out: the Soviet Molniya series reach a distance of 40 000km (25 000 miles). The best situation for a comsat is the geostationary orbit, a circular orbit 35 900km (22 300 miles) out above the equator. Here a satellite completes one orbit in the same time that the Earth turns once, so it appears to be suspended above the same spot. Several dozen satellites now use this orbit.



orbit above the equator. A study in 1979 concluded that the NOAA and GOES satellites are saving the US alone some $170 million per year in their ability to detect sea lanes closed by ice and to increase the accuracy of weather forecasts.

Following their experimental "remote-sensing" Nimbus series (launched 1964–75), the US have launched three Landsat Earth resources satellites (1972, 1975, 1978) to photograph the Earth's land surface. They follow very low orbits, 917 km (570 miles) up, which carry them over the poles. As well as mapping previously unsurveyed parts of the Earth, the Landsats can monitor the development of crops, the spread of drought, and large-scale pollution.

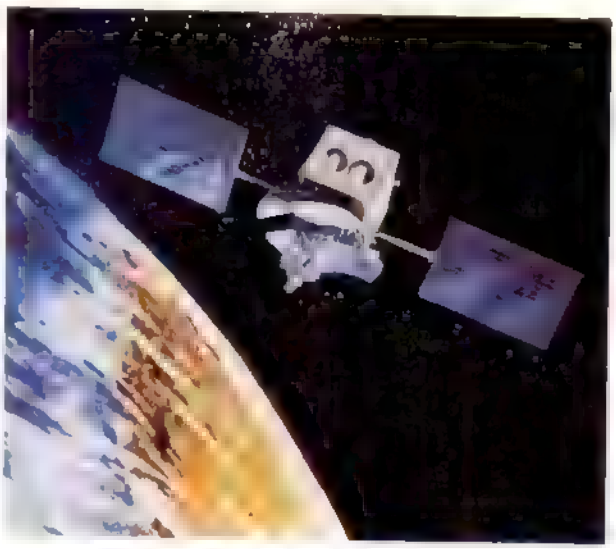
Military satellites

Earth-surveillance is also important to the military. Like Landsat, military satellites can send TV pictures back to Earth, as the US proved as early as 1961 with their Samos 2 satellite. Such messages can however be intercepted, and have limited resolution. Many military satellites instead take photographs on fine-grain film, then eject the film canister to be caught in the air by a high-flying aircraft. The US perfected this technique with the Discoverer series of satellites back in 1960.

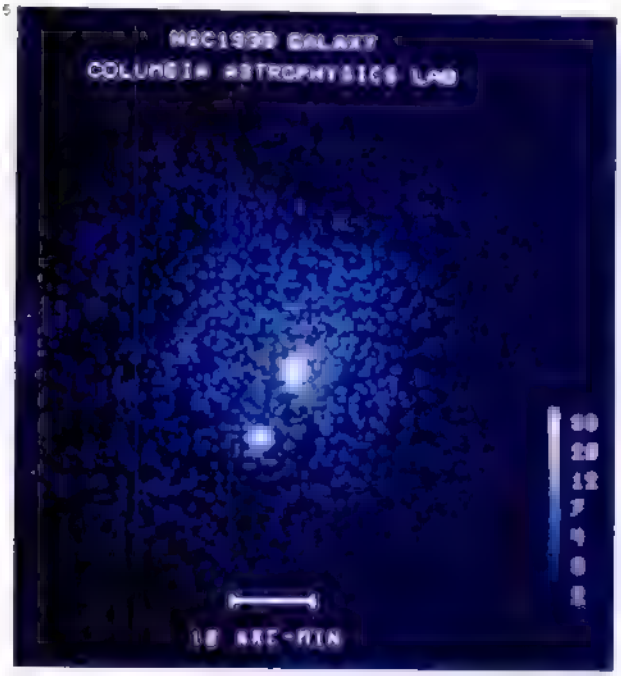
The US and USSR launch military satellites especially to monitor potential trouble-spots: the Iran/Iraq war of 1980, for example, provoked a spate of spy satellite launches. They are put into unusually low orbits to see the finest details possible, and as a result they suffer severe atmospheric drag and re-enter the atmosphere in only a few weeks.

Another military requirement is for precise position-location on Earth. The first of the US navigation satellites, Transit 1B, was launched in 1960. The latest series, NavStar, to be launched in the mid-1980s, will enable anyone with a suitable receiver to pinpoint his position to within 16m (50ft).

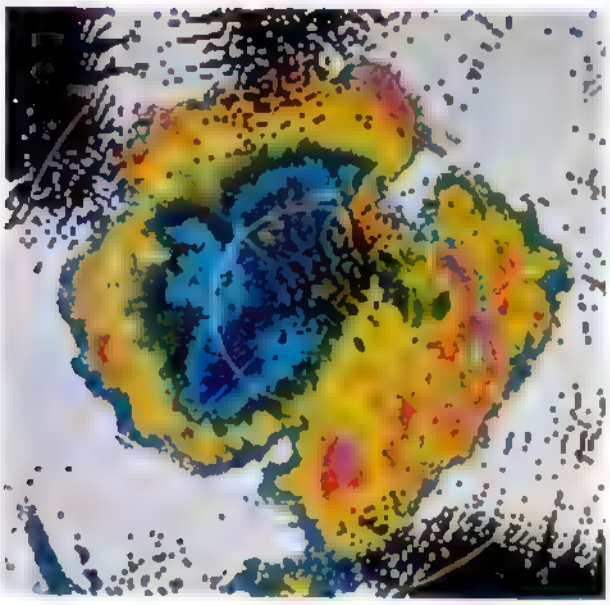
Some satellites look away from the Earth to detect radiations which do not penetrate the atmosphere to reach the ground. Satellites have now explored almost the entire spectrum of radiations from space.



The first commercial US domestic communications satellite, Satcom 1, is owned and operated by RCA. It was launched to provide links with Alaska. Many other companies are now building satellites.



5 Huge clouds of hot gas shine brightly around the galaxy NGC1399 when it is viewed in the short wavelengths of X rays recorded here by the X-ray telescope of the orbiting unmanned Einstein Observatory. Satellites can take instruments into space to observe the universe in ways which are impossible from the ground. X rays from space cannot penetrate the Earth's atmosphere to reach ground-based telescopes, and astronomers did not know the exact shape and structure of X-ray emitting astronomical objects until November 1978 when the Einstein Observatory was launched. The gas shown here is at a temperature of 100 million degrees, and is invisible to ordinary light-gathering telescopes. Other satellites view the invisible universe in gamma rays, ultraviolet and infrared. These observatories have only a limited life: the satellite eventually runs out of the propellant gas used to swing it from observing one source to another.



6 Antarctica and the surrounding ocean are clearly recognizable in this false-colour view based on measurements from the Nimbus 5 satellite. The colours represent the intensity of the microwave radiation from the Earth's surface at a wavelength of 1.55cm. The microwave intensity depends on temperature and the surface's radiating efficiency. The central blue region is snow on the inland plateaux; the surrounding green area the ice sheets covering the rest of the continent. Orange represents pack ice off the coast in the summer; in the winter the Antarctic ice patch is much more extent.

7 The Earth resources satellite Landsat 3, here undergoing pre-flight tests, derives from the early Tiros and Nimbus weather satellites. Launched in March 1978, Landsat 3 follows a polar orbit at a height of 917km (570 miles) that takes it round the Earth once in 103 minutes. As the planet turns beneath it, Landsat passes over a different strip of ground on each orbit; after 251 orbits (18 days) it has surveyed the whole Earth. Landsat 3 has two camera systems. One takes ordinary camera photos, while the other analyses five wavelengths – red and green visible light, and three infrared wavelengths – to distinguish rock, soil, water and healthy and diseased vegetation.

Probing the planets

Without our nearby Moon to practise on, it is difficult to see how we could have come so far in less than two decades of planetary exploration. Space probes initially developed from unmanned lunar probes, such as Ranger, Orbiter and Luna, have now surveyed all the planets known to ancient man and are pushing further beyond. Unmanned craft have made controlled landings upon the surfaces of Earth's two nearest neighbours, Mars and Venus, sending back detailed reports on the conditions there.

First encounters with other worlds were discouraging. The US Mariner 2 space probe, passing within 35,400km (22,000 miles) of Venus on 19 June 1963, found a searing-hot planet swathed in a dense atmosphere of carbon dioxide. Mariner 4 sent back a similarly dismal picture of Mars: the 21 photographs received on Earth on 14 July 1965 showed a barren, cratered world, apparently as dead as the Moon.

Probes to the hell-planet

Subsequent probes have only served to confirm the hostile first impression of Venus. The Soviet Venera 7 (December 1970) and Venera 8 (July 1972) were only able to land on the planet after both craft had been considerably reinforced to withstand the high atmospheric pressure – 90 times that of the Earth. Four earlier Venera craft had been crushed by the extreme pressures before reaching the surface.

The next to visit this hell-planet was the US Mariner 10 which flew within 5,760km (3,580 miles) of its cloud tops in February 1974. It transmitted photographs of Venus's murky atmosphere before flying on to cratered Mercury, which it photographed from only 271km (168 miles). The first and only photographs of Venus's surface came from two Soviet probes, Venera 9 and 10. They landed within days of each other in October 1975, briefly transmitting back photographs of two somewhat dissimilar kinds of terrain before they succumbed to the extreme conditions. In December 1978 Venera 11 and 12 and the US Pioneer Venus craft reached the planet and released smaller probes to investigate the complicated structure of its atmosphere. The Pioneer Venus Orbiter probe spent the following year mapping the planet's hidden surface by radar.

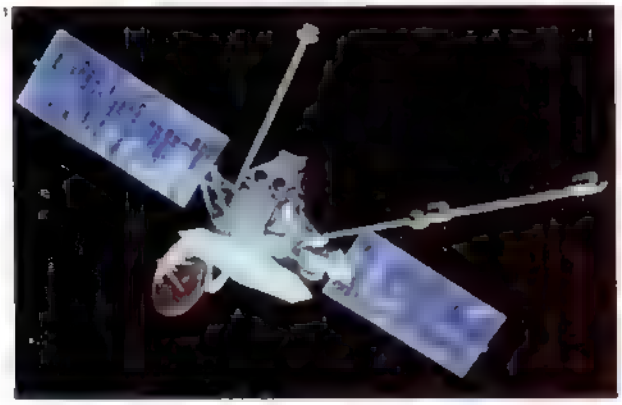
Investigating the deserts of Mars

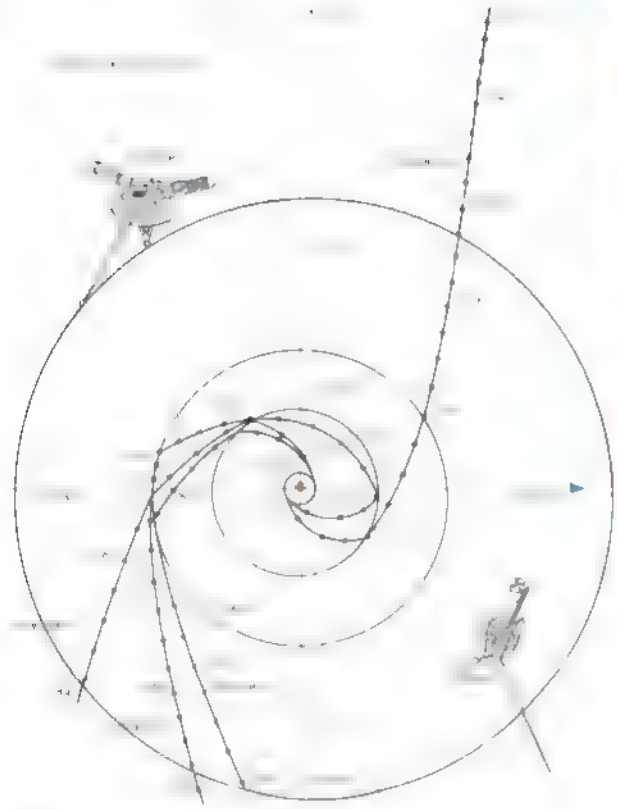
Photographs from the US fly-by probes Mariner 6 and 7 (1969) confirmed the barren, Moon-like quality of the surface of Mars. Then, on 14 November 1970, Mariner 9 went into orbit around the red planet. At first its TV cameras saw practically nothing because Mars was enveloped in a global dust storm. When the dust cleared the view was astonishing: on the hemisphere unobserved by the fly-by probes were huge chasms, towering volcanoes and narrow, dried up river beds. But none of the 7,329 photographs from Mariner 9 showed any direct evidence for the much-hoped-for "life on Mars": a closer look was needed, preferably from the surface. The USSR's attempts to do just this between 1971 and 1973 were unsuccessful.

Meanwhile the US team had planned its own search for life on Mars with the two Viking space probes. Viking 1's arrival on the planet was scheduled to coincide with

CONNECTIONS

See also





1 The Mariner 10 space probe accomplished the first ever [illegible] planet encounter in 1974–75 when it flew past both Venus and Mercury. Solar panels powered the experiments.

2 Four US space probes have now flown into the outermost regions of the Solar System following the trajectories marked here. Pioneer 10 will be the first to leave the Solar System.

3 The two Viking lander craft, which touched down on Mars in 1976, were complicated laboratories designed to test the Martian environment and search for signs of life. Here a Viking lander is tested in the laboratory prior to launch. The three legged craft is over 2m (6 ft) high and 3m (10 ft) across. It weighs over half a tonne. Its 3m (10 ft) long extendable arm is at the rear.

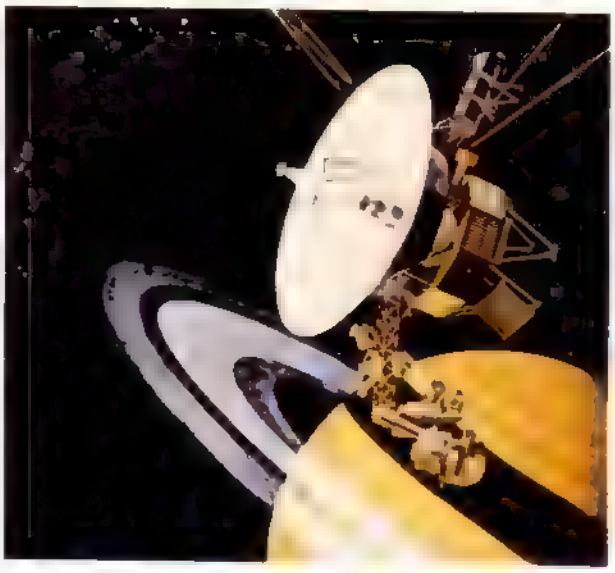
the nation s celebration of its Bicentenary of Independence on 5 July 1976, but Mars did not co-operate Orbiter and lander remained coupled together for two weeks after arriving in Mars orbit while NASA controllers searched desperately for a safe landing site The orbiter continued to take highly detailed photographs of the planet below, while the Viking 1 lander eventually came down safely in the smooth Chryse Planitia region on 20 July Within 25 minutes of touch-down, US scientists were treated to panoramas of the Martian landscape The wastes of Mars looked very like a barren terrestrial desert As well as cameras, Viking 1 and Viking 2 (which landed in the Utopia region on 3 September) carried life-detection experiments, but the result of these, while initially encouraging, in the end turned out to be negative

Exploring the outer Solar System

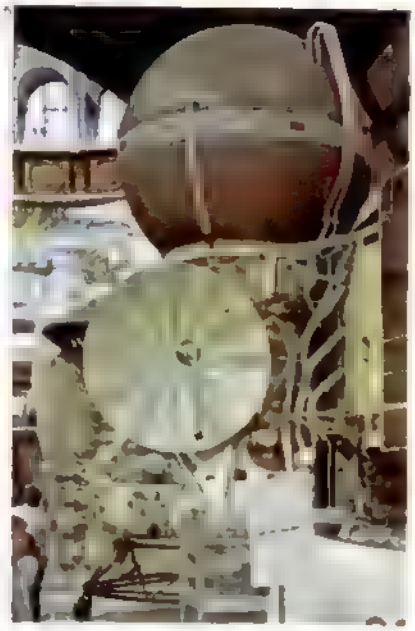
Beyond Mars lies a zone once considered to be a hazard to space probes – the asteroid belt But four US craft bound for Jupiter and Saturn have now passed through this without mishap The relatively simple Pioneer 10 was the first, its experiments powered by a nuclear generator instead of the conventional solar cells, which are useless at so great a distance from the Sun In 1973, Pioneer 10 s camera recorded Jupiter's clouds from 130,300km (81,000 miles), so close that its sensitive instruments were damaged by the planet's intense radiation field

But it is the two US Voyager space probes that have revolutionized our knowledge of the two giant planets. Built on the tried and trusted Mariner design, these vehicles have gathered enough information to keep scientists busy for years Voyager 1, which passed Jupiter in March 1979 and Saturn in November 1980, has now completed its task and will follow the Pioneers out of the Solar System Voyager 2 is bound for Uranus in January 1986 and Neptune in August 1989

Beyond this, the future of planetary exploration is uncertain Yet Oleg Gazenko a Soviet space medicine expert, said recently, It is difficult to give an exact date for a manned flight to Mars .. but I believe it will be before the year 2000



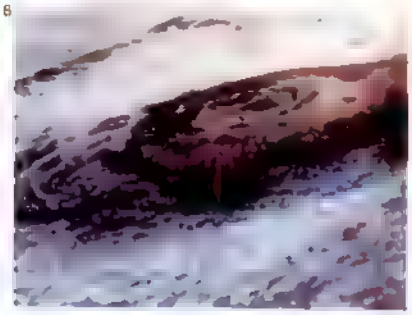
Dominated by its communications dish, Voyager 2 passes Saturn Signals travelling from the distant [illegible] to reach Earth



4 The Voyager 1 space probe flew past Jupiter in March 1979 [illegible] pictures of the giant planet and its major moons In this [illegible] Jupiter is marked by [illegible] Ganymede [illegible] Europa [illegible] and the densely cratered Callisto bottom right

5 The 1m (3 3 ft) diameter Venera capsule (lower right) had to be enormously reinforced to withstand the pressure of 90 Earth atmospheres on Venus's surface

6 Mars's highest volcano, 26km (16 miles) high Olympus Mons is surrounded by clouds in this photograph by the Viking Orbiter



Colonizing space

The 1960s saw the pioneering space flight missions, culminating in the lunar landings in 1969. Subsequently the United States concentrated on reducing expenditure on rockets by developing a re usable launcher the Space Shuttle while the Soviet Union poured its efforts into establishing manned space stations in low orbits around the Earth. Both approaches are essential to the utilization, and eventual colonization, of space by man. By the end of the 1970s, the USSR had achieved great succes es with the Salyut stations while the much delayed American Space Shuttle had yet to fly, the Soviet Salyut 6 space station was being regularly manned by multinational crews and Russian cosmonauts had clocked up almost twice as many hours in space as their US counterpart.

Early Soviet problems

The Soviet Union s programme suffered a severe setback in April 1967 with the death of Vladimir Komarov (1927 67) in the wreckage of Soyuz 1. As Komarov had brought this craft the first of the three man Soyuz design down prematurely from orbit, it had gone into a spin, tangled its parachute and plummeted to Earth.

The next few Soyuz craft were more successful. In January 1969 Soyuz 4 and 5 docked together. Nine months later, three more Soyuz craft practised manoeuvres as a "space fleet". Few outside observers noticed, because in the meantime the American Apollo 11 had landed on the Moon.

The first Soviet space station, Salyut 1 was launched unoccupied on 19 April 1971. Its first occupants the crew of Soyuz 11 stayed in Salyut for a record 23 days. But as Georgi Dobrovolski, Viktor Patasayev and Vladislav Volkov prepared to re enter Earth's atmosphere, a faulty valve vented the atmosphere of Soyuz 11 to space, and the cosmonauts died instantly. Salyut 2 disintegrated before a crew could be launched to it.

Skylab

Meanwhile, the US had prepared its first space station, Skylab. This was not purpose built, but was the upper stage of a Saturn 5 rocket from a cancelled Moon flight, with the empty fuel tanks converted to living quarters and laboratory space. At 75 tonnes, it was four times more massive than the Soviet Salyut.

Skylab, too, had its teething problems. At launch in May 1973, a micrometeoroid shield ripped off, pulling away one of Skylab's two solar panels and jamming the other. The space station began to overheat, and it had no power. Fortunately, the first visiting crew was able to rig up a "parasol", and to extend the remaining solar panel by hand. The three Skylab crews of three astronauts each, set successively longer endurance records of 28, 59 and 84 days.

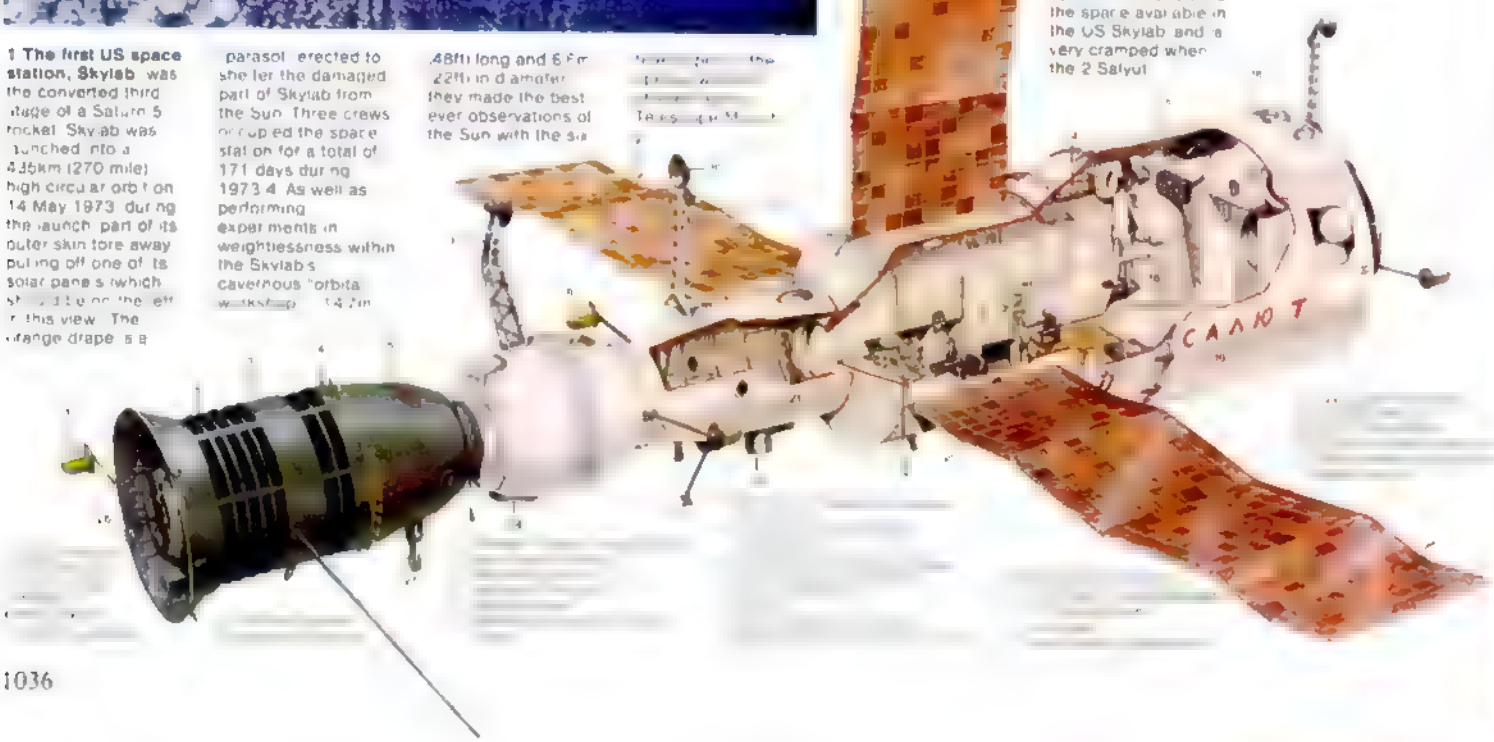
The last Skylab crew returned to Earth in February 1974. US space scientists expected Skylab to stay in orbit for at least nine years, until a Space Shuttle could boost it up to a safe orbit. But they misjudged the atmospheric drag at Skylab s height, and the huge space station re entered Earth's atmosphere out of control in July 1979. Thousands of fragments fell over the Indian Ocean and Western Australia, but caused no serious damage.

CONNECTIONS

See also

1 The first US space station, Skylab was the converted third stage of a Saturn 5 rocket. Skylab was launched into a 435km (270 mile) high circular orbit on 14 May 1973 during the launch part of its outer skin tore away pulling off one of its solar panels (which [illegible] on the left in this view). The orange drape is a parasol erected to shelter the damaged part of Skylab from the Sun. Three crews occupied the space station for a total of 171 days during 1973 4. As well as performing experiments in weightlessness within the Skylab s cavernous "orbital workshop" [illegible] 14.7m (48ft) long and 6.7m (22ft) in diameter they made the best ever observations of the Sun with the six [illegible]

2 The Soviet Salyut 6 is the most successful of the purpose-built Salyut series. It has been inhabited for almost half of its first four years in space and its crews have set successively longer endurance records. In 1979, Vladimir Lyakhov and Valeri Ryumin spent almost 6 months (175 days) in Salyut 6. 8 months later, Ryumin was back, having stood in for an injured cosmonaut. His companion this time was Leonid Popov and they stayed just over six months (185 days). In this period of 20 months, Ryumin had spent 12 in space, to become the world's most travelled man, with a total journey of 240 million km (150 million miles), more than 600 times the distance to the Moon. Salyut 6 weighs 19 tonnes and is 13m 43 ft long overall. It has two docking rings, so that the resident cosmonauts can receive visitors in another Soyuz or supplies from an unmanned Progress transporter. Salyut 6 has a volume of 90 cubic m (3,000 cubic ft), less than one third the space available in the US Skylab and is very cramped when the 2 Salyut cosmonauts play host to a pair of visitors. Among the experiments on board Salyut 6 are furnaces for making ultrapure metal alloys, gallium arsenide materials and semiconductor crystals for electronics. Cosmonauts have also tried to grow vegetables, including dill, parsley and cucumbers. Prolonged weightlessness causes muscular atrophy, loss of body fluids and a gradual weakening of the bones. As counter measures cosmonauts take exercise on a treadmill and a cycle unit, and during their daily activities wear tight elasticated suits [illegible]



Co-operation in space

While Skylab was [illegible] project the Salyut space stations were part of an evolving programme. Salyuts 3, 4 and 5 were launched in June [illegible], December [illegible]74 and June 1976 respectively, both they and the Soyuz ferry craft became more reliable, although Soyuz [illegible] pressure [illegible] for safety which reduced the number of cosmonauts to two in each craft.

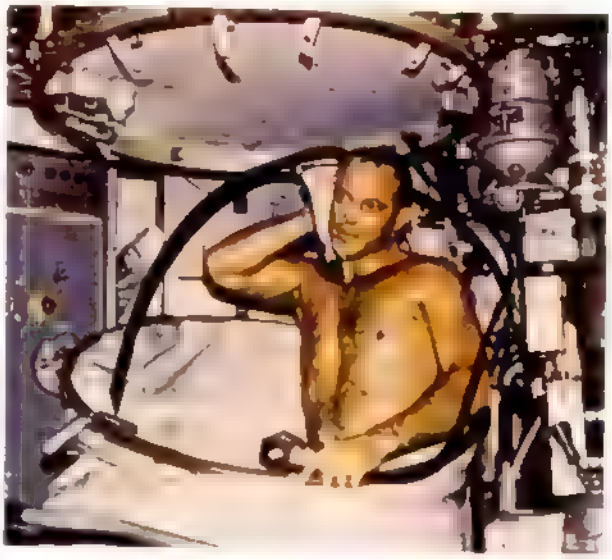
The most memorable event of the mid 1970s was the Apollo-Soyuz Test Project, the first [illegible] joint US-USSR space programme. In July 1975 the US launched its last Apollo craft on a remaining Saturn IB booster. The Apollo carried a [illegible] up with the Soviet Soyuz 19. Astronauts and cosmonauts worked together in the combined spacecraft.

A year in space

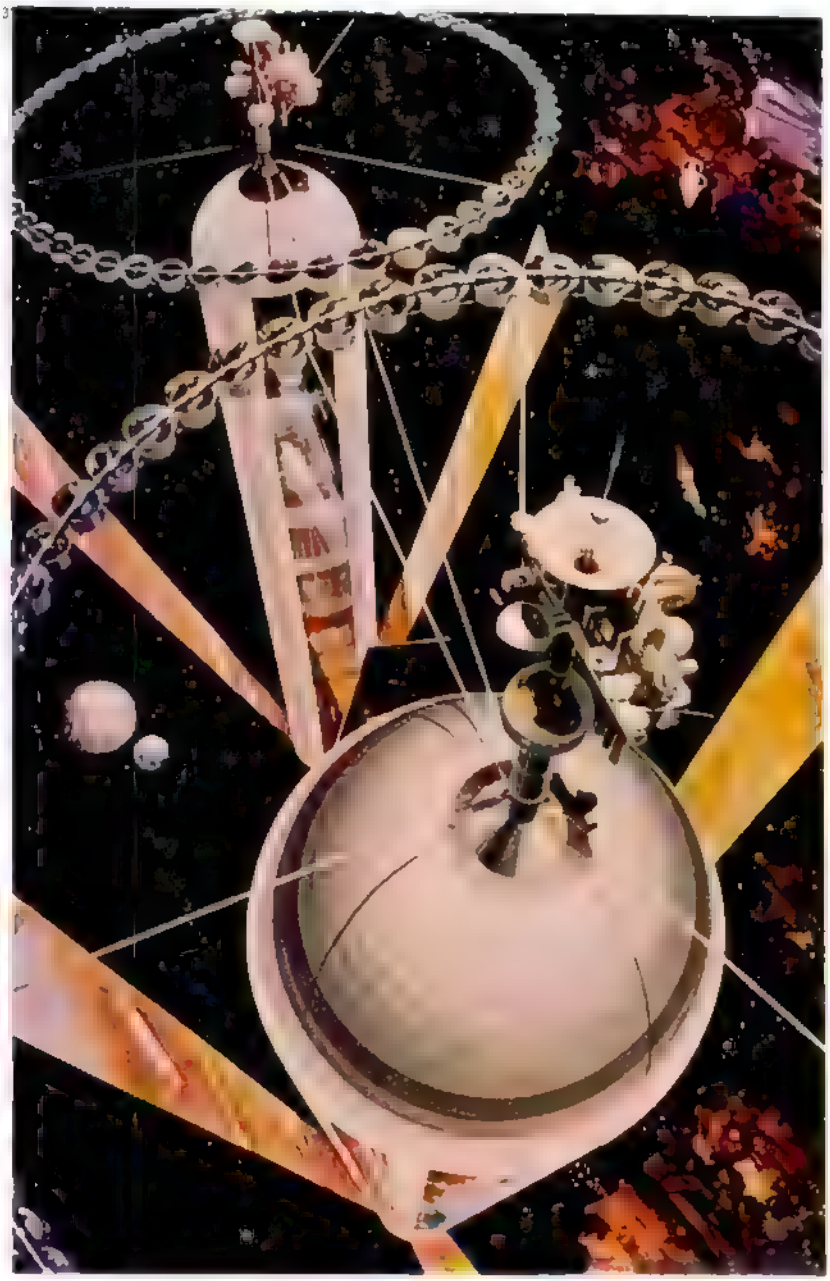
With Salyut 6 launched on 29 September 1977, the Soviet space station programme at last fulfilled its promise. The station's first crew, Yuri Romanenko and Georgi Grechko, was involved in several innovations. The cosmonauts entered Salyut 6 from Soyuz 26 in December 1977, and a month later welcomed Pole to space visitors, the crew of Soyuz 27. Soyuz craft have only a limited orbital life so the visitors departed in Soyuz 26, leaving the fresher craft for the return of Romanenko and Grechko in March 1978. Before this, Salyut 6 was refurbished by the first unmanned supply craft, Progress 1. The Progress craft, an unmanned version of Soyuz, can carry a tonne of fuel and 1.3 tonnes of supplies.

Romanenko and Grechko also received a second human visitation from the first multinational crew. One of the cosmonauts in Soyuz 28, Vladimir Remek, was a Czech.

This highly successful pattern continued. During its first four years, Salyut 6 was visited by six main crews, totalling in all some 23 months in orbit. Thirteen Progress craft kept it stocked. Ten crews came in [illegible] nine other [illegible] from Poland and East Germany, Cuba, Mongolia and Vietnam.



Astronaut Jack Lousma [illegible] S[illegible] space [illegible] has a water [illegible] space in zero gravity shower [illegible] a sealed bag [illegible] the weightless water [illegible] water [illegible] the [illegible]



3 According to some experts, the [illegible] space stations [illegible] and [illegible] here. Each [illegible] (19 miles) long and 6km (4 miles) across. [illegible] retracted into each [illegible] Each colony [illegible] the rapid rate of [illegible] every two minutes so that the [illegible] [illegible]

4 During the joint US-USSR Apollo-Soyuz Test Project [illegible] the Soviet authorities released [illegible] photos of the Moscow Control Centre. Above the [illegible] photo is a world map showing the progress of the mission, and a huge TV screen displaying the [illegible] and astronauts at work together in space. The Control Centre co-ordinates spacecraft [illegible] the USSR's main site at Tyuratam. This cosmodrome has five times the area of Cape Canaveral and as many as 80 launch [illegible]

The Space Shuttle

Twenty years to the day after Yuri Gagarin made the first manned flight into space, the US opened a new era in spaceflight with the lift off of the first re-usable launch vehicle. Space Shuttle *Columbia* blasted off from Cape Canaveral on 12 April 1981, after two days in orbit it landed on an airstrip in California, and was refurbished for a second lift-off later in the year. With a fleet of four Space Shuttles, each flying at least 100 times, the US hopes to cut the cost of space travel by nine tenths.

Birth of the Shuttle

By the end of the 1960s, the American space agency NASA realized it could not continue with the waste and expense of large conventional launchers such as Saturn 5, a $360 million rocket that can be used only once. The project to create a re-usable "shuttle" was authorized by President Nixon in January 1972, the year of the last Apollo Moon flights. The first proposal was a completely re-usable two-stage shuttle, each stage winged like a plane and piloted by two astronauts, but this would have cost double Nixon's budget ceiling of $5,200 million.

Instead, NASA opted for a rather more complicated, but cheaper, shuttle. It has one manned, winged craft, the Orbiter, which carries the main engines and draws fuel from an external tank. To assist the launch, two solid-fuel boosters are strapped to the sides of the external tank. Together, the combination stands 56m (184 ft) tall, and fully refuelled it weighs 2,000 tonnes.

At launch, the three main engines in the Orbiter and the two boosters fire together, creating the most powerful launcher in existence today. They produce a thrust of almost 3,000 tonnes, little short of the record 3,400 tonne thrust of Saturn 5, though the Shuttle is half as tall and two-thirds the weight.

The solid fuel boosters are spent two minutes into the flight, and they are dropped, descending by parachute into the sea to be recovered and re-used. After about eight minutes, the external tank is empty, and the main engines cut out. The tank is dropped and burns up in the atmosphere – it is the only part of Shuttle that is not re-used. Two smaller rocket engines on board the Orbiter push it up to a higher orbit – between 180 and 500km (100 and 300 miles).

After orbiting for between seven days and a month, the 75-tonne Orbiter returns, firing its engines in reverse to slow down its progress, and descends towards the atmosphere. Turning to face forward again, the Orbiter's underside is heated red-hot as it meets air friction. Once in the lower atmosphere, its wings give the Orbiter lift, and it glides to a landing at 350 km/h (220mph).

Three years' delay

The first Shuttle launch in 1981 was three years behind the original target set in 1972 due to two major development problems.

The main engines have to operate at higher pressures and higher pump speeds than any before. They also have to be more reliable, because their re-use means a total firing time of 7½ hours. Components were not tested individually, and during trials unexpected minor failures caused widespread damage to the engine assembly. Eighteen engines had to be built before three were certified as reliable enough for *Columbia*

CONNECTIONS

See also

1 John W. Young (left) and Robert L. Crippen were Commander and Pilot on the first flight of the re-usable Space Shuttle *Columbia*. It was Crippen's first flight, but Young is a veteran astronaut. The launch of *Columbia* on 12 April 1981 marked the beginning of his fifth flight – more than any other US or Soviet spaceman. His previous space experience included the first of the US two-man missions, Gemini 3, and the Apollo 10 dress rehearsal for the Moon landings. He took part in the Gemini 10 and Apollo 16 missions too.

2 The launch of the Space Shuttle *Columbia* on 12 April 1981 opened a new era in spaceflight. All parts of the Shuttle except for the 47m (154 ft) high external fuel tank are re-usable. The winged craft, the Shuttle Orbiter, is 37m (122 ft) long and has a wingspan of 24m (78 ft). It is the only part of the Shuttle to go into orbit, and it coasts back down through the atmosphere to land like a heavy glider. The two 45m (149 ft) solid fuel boosters burn for two minutes at launch, and are then ejected, retrieved from the sea, and refuelled. At lift off the three Orbiter engines drawing liquid oxygen and hydrogen from the external tank produce 600 tonnes of thrust; each booster gives 1,200

3 Five of the six US female astronauts now training for flights in the Space Shuttle. From left to right: physicist Sally Ride, engineer Judy Resnik, surgeon Anna Fisher, scientist Kathryn Sullivan, and another doctor, Margaret Seddon. Next to [illegible] Shannon Lucid.

The other problem was the covering of 31,000 heat-resistant tiles that protect the Shuttle Orbiter on re-entry. These comprised an entirely new concept: previous spacecraft had used a heat shield that dissipated heat by flaking away – a system that is not re-usable. The silica tiles cover about three-quarters of the Orbiter's surface, where temperatures reach 400 to 1,200°C (750 to 2,200°F).

The nose-cap and leading edges of the wings get even hotter, up to 1,340°C (2,450°F), and have more expensive carbon-fibre plastic coverings.

In tests, NASA found that the tiles were so brittle they were splitting and falling off *Columbia*. It took an extra 20 months to strengthen the tiles and re-apply them.

The Shuttle programme

An earlier prototype Shuttle, *Enterprise*, had tested the landing systems in 1977 after it had been lifted into the air atop a Boeing 747. John Young and Robert Crippen, the crew on *Columbia's* first orbital flight in 1981, however tested the main engines and heat-resistant tiles, and found no problems.

After four proving flights, *Columbia* will take commercial cargoes into space. It will be joined by three sister-ships – *Challenger* (1982), *Discovery* (1983) and *Atlantis* (1985). Each can lift 29 tonnes into low orbit in its large cargo bay.

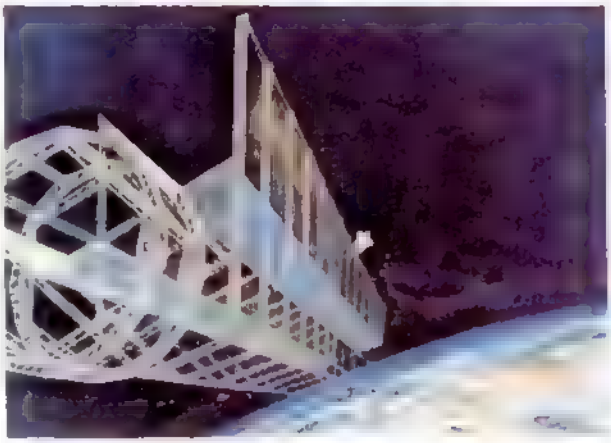
The Shuttle can perform very flexible missions. It can lift satellites – one of the most unusual being the Space Telescope – from its cargo bay to place them in near-Earth orbit. It can also catch satellites to return them to Earth for repair.

An extra booster can be attached to a satellite to push it to orbits higher than the Shuttle can reach, especially the important geostationary orbit 35,900km (22,300 miles) high. The European Spacelab, a laboratory for observing Earth and Sun and investigating weightless conditions, will remain in the Shuttle's cargo bay.

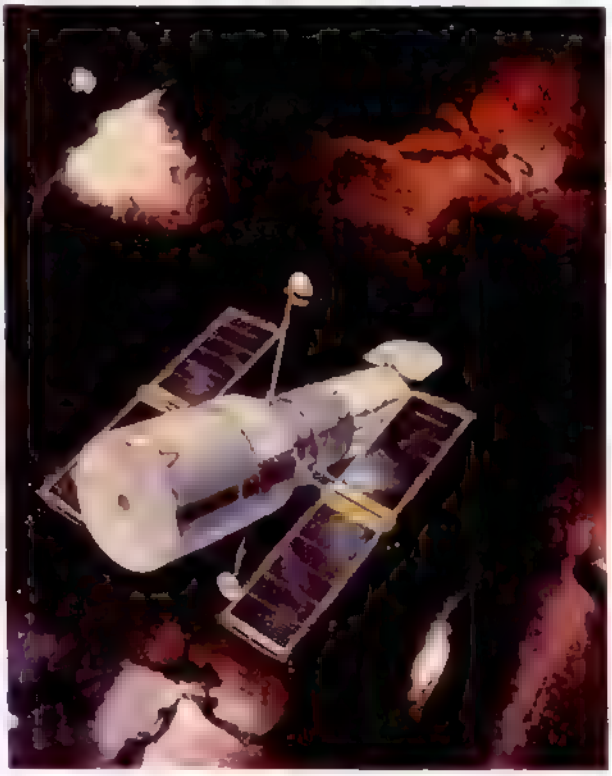
Even before *Columbia's* first flight, the Shuttle was booked up for its first 70 flights. NASA estimates that it can put a satellite into a low orbit for less than a quarter the cost of a conventional launcher.

The first re-usable launcher, the US Space Shuttle *Columbia* lands on the Rogers Dry Lake dirt runway of the Edwards Air Force base in California after its maiden voyage to space. Piloted by John Young and Robert Crippen, *Columbia* spent 2 days 6 hours 22 minutes completing 36 orbits of the Earth before it came down in this flawless landing on 14 April 1981. The flight was trouble-free apart from the loss of 16 heat-resistant tiles fortunately from the top parts which bear only minor heating on re-entry. Edwards AFB was chosen for the test landing because the runway is surrounded by desert which gives a safe landing area. Future shuttles will land in Florida.

4 Launching satellites will be one of the Space Shuttle's main tasks. It lifts the satellite from the cargo bay and places it in orbit with a robot arm 15m (50 ft) long. A Shuttle can lift a 29-tonne payload to the low orbits – up to 500km (300 miles) – that it can reach itself. Satellites intended for higher orbits will carry an extra rocket engine to boost themselves up further. One third of all Shuttle-launched satellites will be for military purposes.



5 The Space Telescope is a large reflector that will be put into orbit by the Space Shuttle in the mid-1980s. It will be unmanned, under remote control from Johns Hopkins University, Baltimore, but Shuttle astronauts will visit the Space Telescope to carry out repairs and change equipment every 2½ years. From its position above the Earth's shifting atmosphere, the 2.5m (94 in) telescope will have a view unrivalled by any of the larger telescopes on Earth.



6 A huge solar power satellite could be the answer to future energy needs on Earth. An array of solar cells hung along giant steel girders several kilometres long – the Shuttle here gives an idea of scale – could convert sunlight into electricity, and the power could then be beamed down to Earth in the form of microwaves. Large receiving arrays on the ground would pick up the microwaves and convert them back to electricity to feed into the national grid. One solar power satellite with an area of 100 sq km (38 sq miles) could provide 5,000 megawatts, enough to supply a city such as London. Although such satellites would provide totally renewable non-polluting energy, the cost would be so high that it is unlikely that any will be constructed. The US government has studied a scheme to construct 60 solar power satellites to provide 30 per cent of its power by 2040 and reckons the cost to be over $3,000,000 million.

Stations in space

The idea of a space station or artificial manned satellite permanently circling the Earth outside its atmosphere was discussed by a Russian scientist Konstantin E. Tsiolkovsky (1857–1935) [Key] about the turn of the century, although at the time it could be regarded as little more than a fantasy. He looked upon the space station both as a stepping stone refuelling base for spaceships visiting other planets and as a laboratory in which scientists could carry out experiments that would be impossible to perform on Earth. He even proposed growing plants in space stations to provide their crews with alternative sources of food and oxygen.

Overcoming gravity

The effect that weightlessness would have on space travellers was unknown. In orbit, where gravitational attraction is opposed by equal and opposite inertia forces, a body experiences no mechanical stress, and astronauts and any loose objects float weightlessly. The same is true of an unpowered spaceship moving in frictionless space towards or away from the Earth.

It was widely believed that even short periods of weightlessness (or zero gravity) would have ill effects on space travellers and thoughts turned to creating artificial gravity. One of Tsiolkovsky's first designs showed a huge cylindrical space station that spun on its central axis. The crew had their feet firmly planted on the inside walls with their heads pointed at the spin axis by action of centrifugal force. Vegetation in a "cosmic garden" grew inwards towards the centre.

As late as 1952 **Wernher von Braun** (1912–77), who worked on the V2 at Peenemunde and who was mainly responsible for the rockets that launched America's first artificial satellite and the Apollo moon craft, proposed a space station that had the form of a huge rotating wheel [2]. The crew quarters were in the rim of the wheel, docking ports for visiting spacecraft were in the centre (hub) and tubular "spokes" allowed people to move from one part of the space station to the other.

Yuri Gagarin's flight in Vostok 1 in 1961 showed that weightlessness is not uncomfortable and since then men have remained in orbit under conditions of zero gravity for nearly three months. However, it is still uncertain whether, after spending much longer periods in space, the human body can adjust to normal gravity conditions without ill effects and we may yet see attempts to create artificial gravity in large-scale vehicles which would enable astronauts to spend possibly limitless periods in outer space.

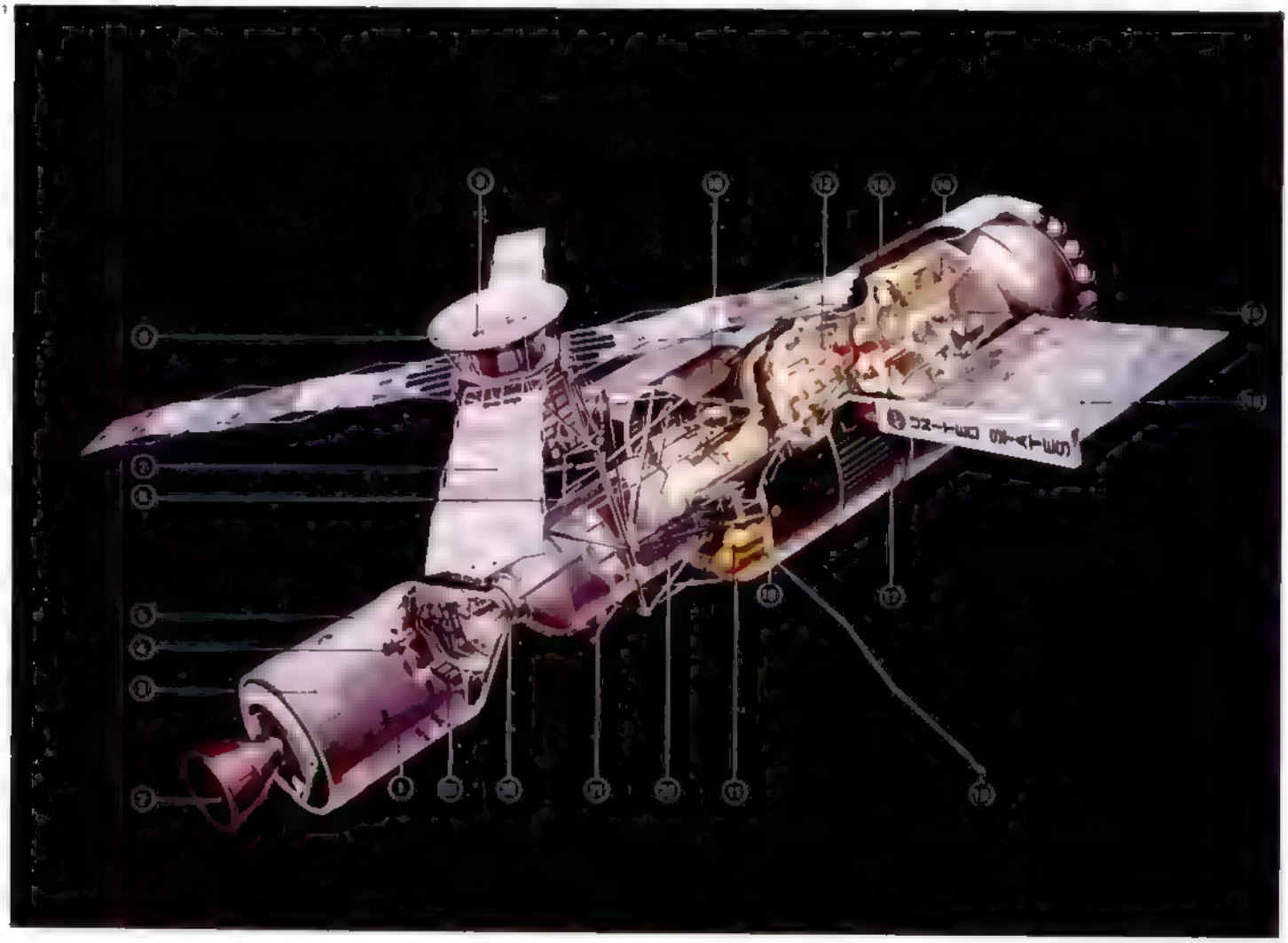
Present and future space stations

True orbital stations were first launched during the early 1970s after experience had been gained with Soviet Soyuz and American Apollo spacecraft, which carried a variety of scientific instruments.

Although the Russians had considerable trouble with the first of their 19-tonne Salyut stations, they have since flown a number of highly successful missions. The Americans had problems too. Skylab was damaged when it was launched and had to be repaired in orbit before space teams could begin their experiments. It was manned by three successive crews who spent 28, 59 and 84 days in space respectively [1].

CONNECTIONS

See also



1 Skylab, launched in 1973 and manned by three crews for a total of 171 days, re-entered the Earth's atmosphere over Australia in July 1979. Its weight was about 82 tonnes and it measured 25m (82ft) long and 6.7m (22ft) across the workshop section. While in orbit, the instruments and systems were powered by solar cells. The various components are indicated by the following key: 1 modified Apollo spacecraft (command module plus service module) to take the crews to the space station; 2 service propulsion system engine with a thrust of 9 100kg (20 000lb); 3 radiators; 4 attitude control jets used in docking, each nozzle with a thrust of 45kg (100lb); 5 crew station in command module; 6 Apollo telescope mount; 7 solar cells, converting sunlight to electricity to power the Apollo telescope mount; 8 sun shield (this gave trouble initially but was later rectified); 9 telescope apertures; 10 oxygen tank and 11 nitrogen tank for the two-gas atmosphere inside Skylab; 12 manoeuvring unit; 13 lower body negative pressure device; 14 gravity substitute workbench; 15 food provisions; 16 solar cells (the foreground wing was torn off during launch); 17 sleep restraints; 18 water containers; 19 aerial; 20 multiple docking adapter; 21 alternative docking port; 22 atmosphere interchange duct; and 23 descent battery packs charged by solar energy from the telescope mount.

The first ferry craft to bring crew and cargo to space stations were versions of existing spacecraft launched by expendable rockets. To reduce the cost American scientists are now building the re-usable "space shuttle", which will take off vertically like a rocket, launch satellites or visit a space station and fly back to Earth like an aeroplane [3]. The winged orbiter is intended to be re-used at least 100 times.

When not used for other purposes, the shuttle can become a miniature space station in its own right. Within its 18.3×4.6m (60×15ft) cargo bay will fit the European Space Lab in which up to four scientists can work without spacesuits for up to 30 days as the spaceplane swings round the Earth. After a mission Space Lab can be replaced by another cargo ready for another flight.

The uses of orbiting stations

The value of orbiting stations is to carry out work that cannot be done more cheaply by unmanned satellites. The first space stations were used for research in biology, chemistry, physics, observations of the Earth's natural resources and studies of the Sun and other stellar objects.

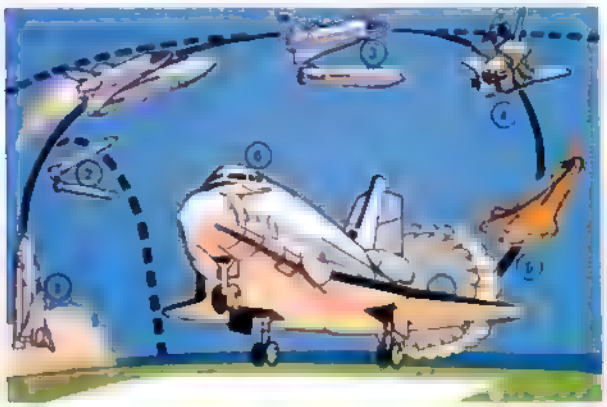
Skylab and the Apollo spacecraft that took part in the "space handshake" with the Soviet Soyuz in 1975 each carried small electric furnaces to melt various metal samples under weightless conditions. In the future it may be possible to make ultra-lightweight foamed steels with many of the properties of solid steel, to combine dissimilar materials such as steel and glass and to grow crystals of great purity for the electronics industry.

The gravity-free environment should also be ideal for isolating biological materials for the treatment of certain diseases and to purify vaccines.

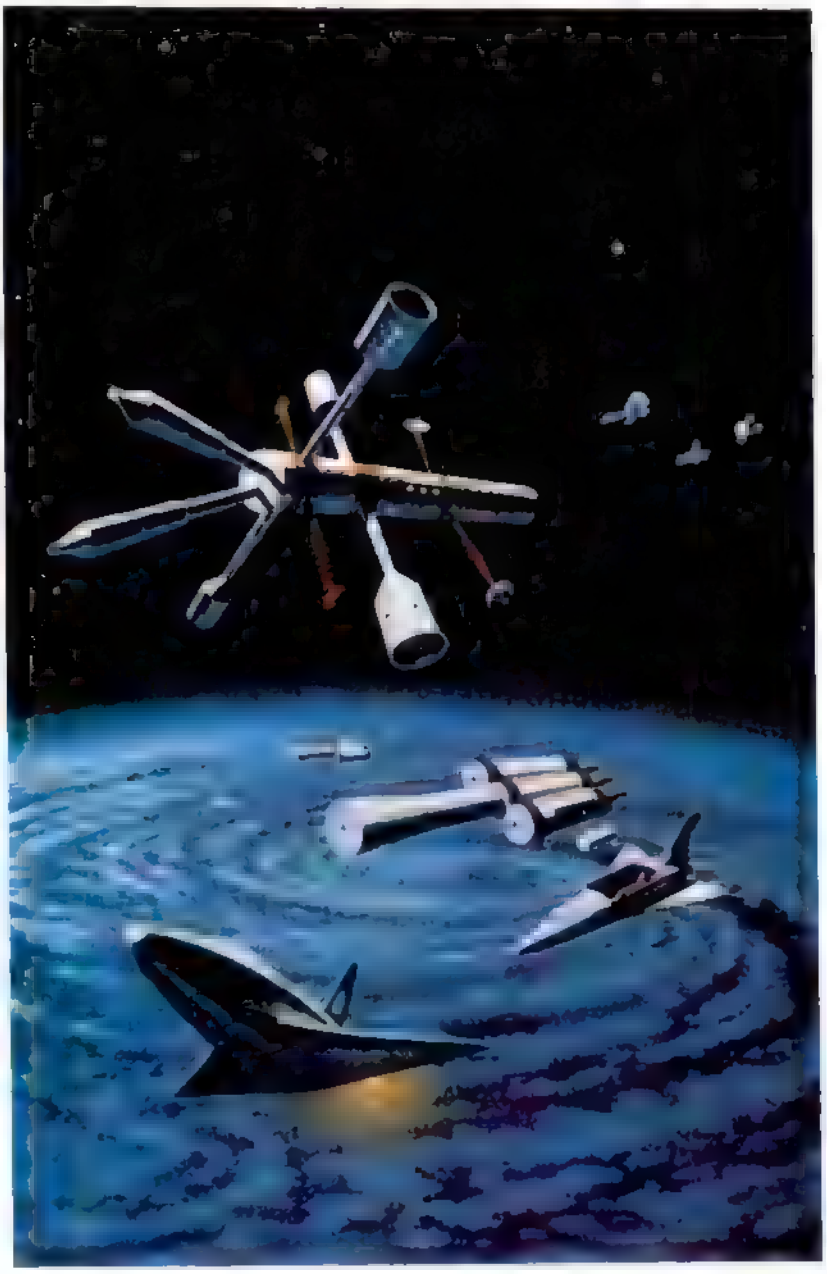
Future space stations will be assembled from modules (units) ferried into orbit by the shuttle. Eventually, it may be possible to build orbiting power stations that have immense solar-cell collectors generating electricity from sunlight. The energy would be beamed to Earth by microwave antennas for use in factories and homes, and would provide a limitless supply of power.

Konstantin Eduardovich Tsiolkovsky was the "father of astronautics" and the man who laid down the principles of space stations. Tsiolkovsky was a shy, deaf Russian teacher whose first papers on astronautics, although written about 1897, appeared in 1902. Even then they aroused little comment and Tsiolkovsky did not become famous until near the end of his life. Today there is a Tsiolkovsky Museum in Kaluga, where he lived. Although he was not an experimenter many of his theories were correct and as a space pioneer he was decades ahead of his time. He was also the first to stress the importance of liquid propellants in rockets.

2 Early space-station design took the form of a wheel with the power supplies in the hub and the crew quarters arranged round the rim. This design was originally worked out to provide for "substitute gravity" by rotation of the wheel. It was then believed that zero gravity might be harmful to astronauts even over short periods, but after Yuri Gagarin's flight and subsequent experiences with Skylab and the Soviet Salyut stations this was disproved.



3 The NASA Space Shuttle has a winged Orbiter that lifts off [1] with a large external tank containing its main propellants and two solid rocket boosters which separate [2] when the craft has climbed 45km (28 miles). The tank is discarded [3] just before the Orbiter goes into orbit. Typical payloads are the European Space Lab, space probes for release in orbit [4], or modules for assembly into a space station. The Orbiter uses its manoeuvring engines as retro-rockets to re-enter the atmosphere [5], when it endures high temperatures, for which it is reinforced with heat-resistant materials. It lands back at base [6].



4 Future orbiting space stations will be different from the neat wheel design of pioneer days. A station assembled in space is designed to move round the Earth at a height of about 500km (300 miles), and will accommodate a crew of up to 100 members. In this picture we see a space shuttle that has just delivered a propellant module for a spaceship bound for Mars. The nearer shuttle has just fired its retro-rockets to start its journey back to Earth. Below the surface of the Earth itself is largely covered with cloud. Various other vehicles are shown in the black sky.

Colonizing the Moon

The Apollo missions to the Moon were essentially reconnaissances [Key]. All that Apollo could hope to do was to take three men to the neighbourhood of the Moon, land two of them on the surface for a brief period and then bring all members of the crew home safely. There was no provision for rescue in the event of a failure of the lunar module during the surface expedition and the time scale was very limited. Yet Apollo was an essential part of the main programme of lunar colonization and it showed that there is no reason why bases should not be set up on the Moon sometime in the future.

Problems encountered on the Moon

There is no question of turning the Moon into a kind of second Earth. The main problem is the lack of atmosphere. Unfortunately, the Moon is an airless world and there is not the slightest chance of providing it with a breathable atmosphere; the low escape velocity means that it is incapable of retaining a dense atmosphere similar to that of the Earth. Lack of atmosphere means a total lack of water and it now seems that – contrary to earlier expectation – it will not be practicable to extract water from the lunar rocks for the simple reason that there is none to extract. Neither is there any hope of finding underground supplies of ice. Colonists of the future will have to take everything with them and it will be a long time before a lunar station can hope to become self-supporting.

The development of lunar bases

By 1990 or thereabouts, the first permanent lunar bases should have been developed and these will be a great advance on the chemically propelled vehicles of Apollo. By then too, the space-station projects should be well under way and it will be practicable to consider going back to the Moon. Possibly the first step will be to send supplies to the surface, setting them down at a prearranged point to await the arrival of the explorers so that when the astronauts land they will find supplies of various kinds ready for use.

The lunar modules themselves may be used as the first bases, but this pioneer phase should not last for long and more elaborate designs should be developed quite quickly. One pattern, dating from the 1930s, is that of a series of domes, each kept inflated by the pressure of air inside it and equipped with a system of airlocks for the exit and entry of the crew members. This kind of design might be developed. Fortunately, it is now known that there is no danger from meteoritic bombardment so that the relative fragility of the domes should not prove a major problem. In the pre-Apollo days it was thought that it might be necessary to construct lunar bases underground for protection.

Even when space-shuttles have been perfected, the expense of travelling between the Earth and the Moon will still be considerable and every possible method will have to be used to cut supply journeys to a minimum. Everything (including human waste products) will have to be "recycled", particularly the atmosphere. The colonists will spend long periods on the Moon which involves making the conditions as comfortable as possible; inside the base it will be essential for the colonist to be able to take off his spacesuit and behave as naturally as is possible under conditions where gravity is

See also
The Moon

1 Science-fiction writers have always been attracted by the Moon. Jules Verne (1828-1905) described a circumlunar voyage more than 100 years ago and H. G. Wells (1866-1946) a fantastic world peopled by insect-like beings. In 1902 came the first famous Moon voyage film, produced by George Méliès (1861-1938). This frame from the film shows the arrival of the rocket (to the distress of the Moon) after which the space travelers go for a walk on the surface, not forgetting to put up their umbrellas to shield themselves from the strong sunlight. The film was short but popular and a favourite at scientific conferences.

2 The lunar Mare Imbrium [A] and the Apennine mountain range [B] was the site of the Apollo 15 landing and it is not inconceivable that it will also be the site of an early lunar base. It lies well away from the Moon's equator. The Mare Imbrium is one of the more level parts of the Moon and the detailed information now available, including samples of rock soil for laboratory examination, confirms that it is a complicated patchwork of lava flows.

3 The first lunar base is unlikely to be at all like the final elaborate stations. Here is the scene after the pioneer expedition has established itself. In the foreground is the basic space-station module which will serve as the centre of the future station. It can accommodate a crew of up to a dozen members and it can also provide all that is needed for a prolonged stay, although in an emergency the expedition can be halted and the crew returned to Earth by means of ferry vehicles. Also shown is a lunar Rover, similar to the vehicles used by the last three Apollo expeditions, which proved to be so successful, a cargo landing craft with a separate cargo module and a lunar drill.

only 17 per cent of the Earth's. There is also the question of recreation, which will be important in view of the relatively long tours of duty that are to be expected. Books, films and musical recordings present no problems but what about physical recreation? No doubt various new sporting activities will be developed on the Moon, suited to conditions of low gravity.

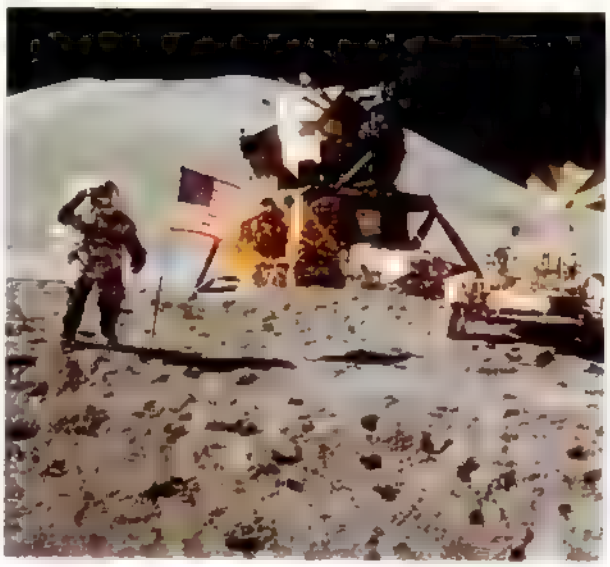
Essential supplies of food

To send all food supplies from the Earth will be impracticable and efforts will be made to persuade edible plants to grow on the Moon. Of course this cannot be done in the open, but inside the domes the principles of hydroponic farming, whereby plants can be grown without soil, may be used. The plants are suspended in netting inside a tank and are fed by liquid nutrients circulated beneath them. The principle has been tested and excellent results have been achieved, so that there seems no reason why it should not work on the Moon.

The early lunar bases may be staffed entirely by scientists. There may be physicists, anxious to take advantage of the low gravity, the limitless hard vacuum and the chance to study all the radiations coming in from space; astronomers, thankful to escape from the restrictions imposed by the screening layers in the atmosphere of the Earth; chemists, biologists, medical men – in fact, scientists of all disciplines. The lunar base should add appreciably to the sum total of our knowledge.

This will be the second phase. Then, as the colony becomes more and more self-supporting, it may be able to take in non-scientists as well, at least for brief visits. The idea of "holidays on the Moon" may no longer be fantastic in 100 years' time. By then, there may be children who have been born on the Moon and who regard it and not the Earth as their home world. Before the end of the twenty-first century the Moon should support not one base, but many, used for a variety of purposes.

It is remotely possible that by the end of the twenty-first century, the indigenous population of the Moon may be demanding independence from the home planet Earth.



Astronaut James Irwin of the Apollo 15 mission is standing in the Hadley Apennines region, one of the possible sites for the first lunar base. The peak of Hadley Delta is seen beyond.

4 A permanent lunar base will probably be set up in the far north of the Moon in order to avoid the intense daytime heat of lower latitudes. (The night temperature is virtually the same at all points on the surface.) The relatively low altitude of the Earth in the sky indicates the high altitude of the site. At the limb of the Earth-turned hemisphere, an observer on the Moon would see the Earth at the horizon (with slight variations due to the irregular orbit of the Moon) while from the Moon's far side the Earth would never rise at all. In this illustration the Earth is shown as full, with its surface detail masked by cloud. It has just passed through the Milky Way into the constellation of Gemini, the Twins. The red star to the right is the semi-regular variable Eta Geminorum. The base is made up of several domes, each with its separate system of airlocks. It will be essential to conserve atmosphere to the greatest possible extent and individual airlocks are necessary to guard against the sudden failure of pressure in one of the domes. Radio aerials and various kinds of instruments are also shown. The illumination is purely by Earthlight (because the Earth is full the Sun is below the horizon to a terrestrial observer, the Moon is new) and the radiance on the lunar rocks is bluish.

The Martian base

In the early part of the twentieth century it was generally supposed that Mars might well be able to support terrestrial life, and that it might even be inhabited. In 1877 the Italian astronomer Giovanni Schiaparelli (1835–1910) had set the world talking about intelligent beings on Mars who had constructed a vast network of canals to irrigate their dying planet. "All the vast extent of the Continent," he wrote, "is furrowed upon every side by a network of numerous lines of a more or less pronounced colour . . . Some of the shorter ones do not attain 300 miles, others extend for many thousands." It did not seem to matter that, at the distance of Mars, such features would have to be tens of kilometres across to be seen at all.

Life on Mars

The myths were shattered once and for all when the first space probes reached Mars in the 1960s and 1970s. Instead of great Sahara-like deserts, there were thousands of Moon-like craters, huge volcanoes and immense canyons and features such as the dried-out beds of ancient rivers (or thin lava streams according to one theory). Whatever water exists on Mars may be locked up beneath the surface in the form of ice and permafrost. The glistening white polar caps appear to be mainly water ice. The climate is cold and the thin, mainly carbon dioxide atmosphere (5 to 8 millibars pressure) does not allow water to exist in a free state.

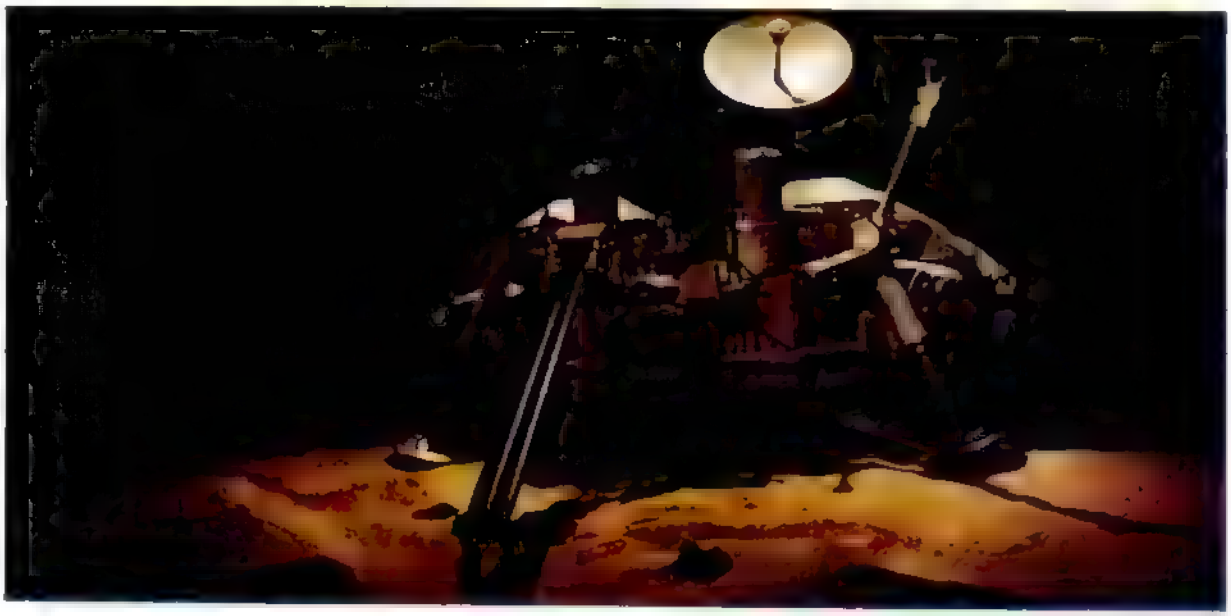
The space probes solved another mystery. The so-called "dark areas" on Mars which some astronomers had linked with the growth and decay of vegetation, seem to have been caused by wind storms which redistributed light-coloured dust over a darker surface. In 1975 two Viking spacecraft, each carrying an automatic laboratory, were sent to see if the soil contained microbes. The quest for life unique to another planet will ensure that exploration continues.

The main problems in sending people to Mars are time and distance. Whereas astronauts can land on the Moon and return in less than two weeks, a Mars ship must leave the Earth–Moon system and embark on an immense journey which takes it right around the Sun. Before such a project can even begin it is necessary to build a space station in Earth-orbit which can be used as an assembly and refuelling depot. The men and women who make such journeys must learn to live and work in space for 1½ to 2½ years.

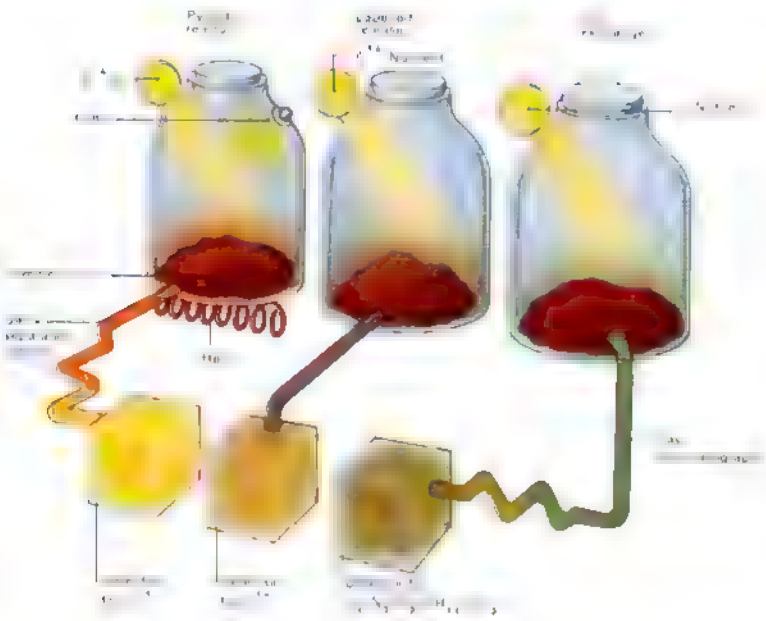
Landing on Mars

Opportunities to launch to Mars come round at intervals of about 25 to 26 months when Mars is at opposition. One American plan, shelved for the time being because of the enormous cost, involved a possible expedition this century. Two atomic-powered ships each 82.3m (270ft) long were each to carry six explorers. Work had already started on Nerva rocket engines which use nuclear heat to expand liquid hydrogen fuel into a powerful propelling jet. The journey was to take the explorers around the Sun to a point in space where Mars would be in nine months' time. For most of the flight the two ships would be docked together nose-to-nose, separating before they arrived. They were to orbit Mars for 80 days while three explorers from each ship descended to the surface in landing craft.

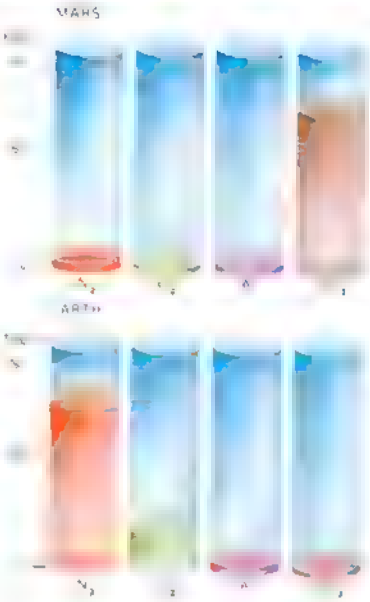
CONNECTIONS



1 The engineering replica (simulator) of the Viking lander at the Jet Propulsion Laboratory (JPL) at Pasadena, California, was used by scientists to solve problems as they occurred with the actual spacecraft on Mars. They were thus able, for example, to free a locking pin on the Viking 1 soil scoop that had failed to eject. JPL commanded the scoop arm to shake out the pin after working out a series of movements with the simulator. The "repair" was effected across more than 340 million km (212 million miles) of space. The sampler scoop, cameras (upper left and centre) and meteorology beam (upper right) can clearly be seen in this photograph.



2 The search for life was an important task for Viking. Each lander has an automatic biology laboratory to which soil samples were delivered by a mechanical scoop in three experiments. Pyrolytic Release was designed to detect any micro-organisms that live by photosynthesis, taking CO_2 and using it to make organic matter in artificial sunlight (lamp). Labelled Release looked for signs of metabolism, evidence that organisms in the soil were maintaining and reproducing themselves. Gas Exchange looked for any exchange of gases between microbes in the soil and atmosphere of the chamber.



3 The main constituents of the Martian atmosphere are here compared with those of Earth's atmosphere.

Inside their pressurized, temperature-controlled landers, the astronauts would enjoy a "shirtsleeve" environment. They would have to don spacesuits before leaving the ship in order to set out their instruments, take samples and make excursions in a Martian roving vehicle. Then they would stow their samples and blast off from Mars, leaving the lower section of the landing craft behind, to rendezvous and dock with their orbiting mother ships. At a pre-calculated time, the ships were to blast out of Mars orbit and continue their long journey round the Sun, flying close to Venus and using its gravity to slow down and arrive back at the Earth station after a round trip of 21 months.

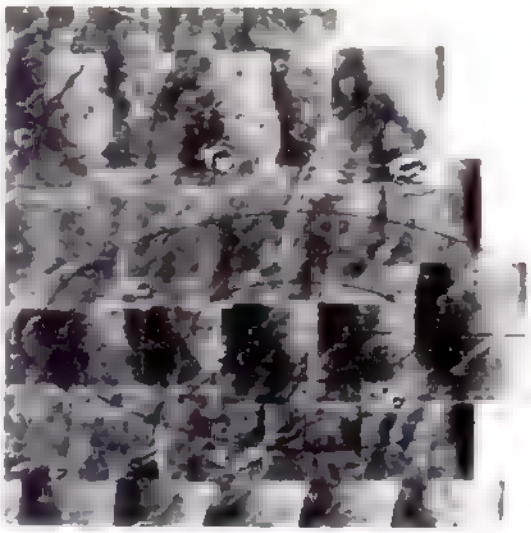
Such an ambitious plan would cost well over $100,000 million, and the US will certainly not attempt it this century. But the USSR apparently has plans to put a man on Mars. Soviet scientists have been quoted as saying they expect to see a cosmonaut in orbit around Mars, if not on the surface, by the end of the century.

Cosmonauts have set ever-longer records for endurance in space – up to six months, which is approaching the length of a Mars trip. An improved Salyut type space station, equipped with a suitable rocket engine, would make an acceptable Mars-cruiser at comparatively small expense.

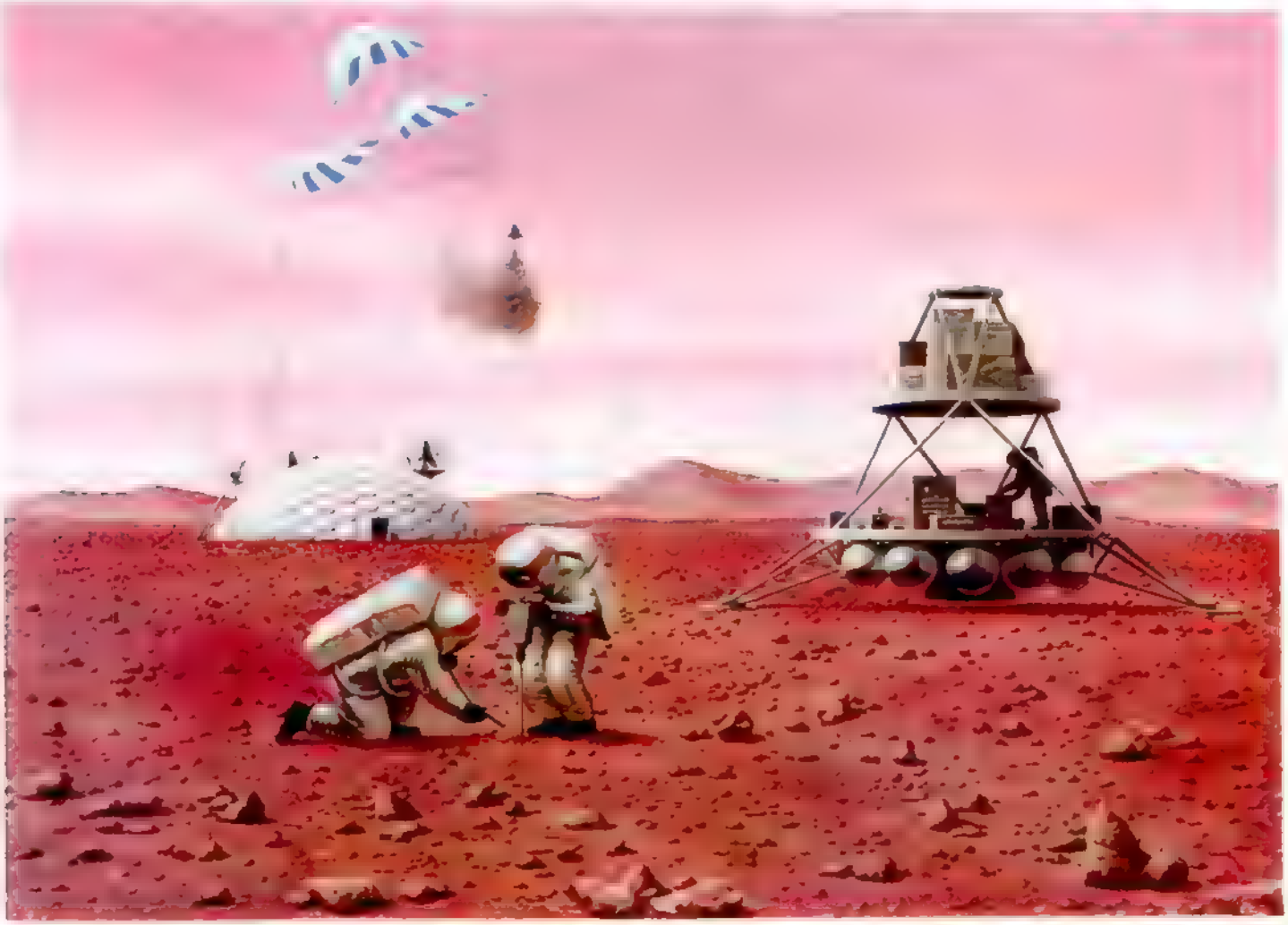
Intelligent machines

A scientific base on Mars would need a much larger investment and its value must be weighed against the ability of intelligent machines to gather scientific information. For example, computer-controlled roving vehicles have been devised to avoid the time-lag in radio communications (up to 23 minutes in the case of Mars) which rules out direct radio-control steering from Earth. The Russians have experimented with a six-legged vehicle with a computer "brain" and a laser "eye".

It is even possible that unmanned spacecraft can be made to land on Mars, release an automatic rover and obtain soil and rock samples at different sites. After the samples have been placed into the return rocket, at the appropriate time it would be launched back to Earth.



The Viking 1 landing site in Chryse Planitia – a channelled lowland some 2.5km (1.5 miles) below Mars' mean "sea level" – lies northeast of the great Martian volcanoes and the Tharsis Plateau. The craft landed near the centre of the target ellipse in the picture.



4 The procedure for establishing a base on Mars will differ from that on the Moon. It will be necessary to build a complete base immediately since Earth is too far away in case of emergency. In this illustration a Mars Excursion Module (MEM) similar to the type needed to carry astronauts to the surface of the planet has just arrived. Its cargo includes an inflatable pressure dome like the one already set up in the background. Another MEM is about to and a roving laboratory which the astronauts will use to explore the ground. The landing technique follows the practice of Viking. After release from the orbiting mothercraft, the MEM fires a braking rocket to descend. It enters the Martian atmosphere at about 16,000km/h (10,000mph) to be slowed first by aerodynamic drag on the blunt heat shield, then by parachutes. A few thousand metres up the parachutes are discarded and retrorockets cushion the touchdown. Equipment on Mars must be prepared for the climate. Pressure domes are double-walled to protect the people inside from the cold. Radar dishes are mounted in radomes mainly for protection from wind-blown dust. Other supplies include nuclear generators for power supply and reserves of oxygen for life-support systems, food and water. Sources of water and oxygen must be established to make the base less dependent on Earth. The search will already have begun with projectile-like probes launched from orbit which penetrate many metres into the soil to determine its chemical composition and the presence of permafrost. In likely places astronauts will drive out in their roving laboratory to make a thorough study. In the picture, geologists are shown taking core samples for analysis.

Exploring the inner planets

Although Venus approaches the Earth more closely than any other planet, it took a long time to discover the true nature of its surface because it is perpetually shrouded by a cloudy white atmosphere. Only slightly smaller than the Earth, Venus orbits some 42 million kilometres (26 million miles) nearer the Sun and intercepts about twice as much heat and light. These simple facts led astronomers to some quite bizzare theories. The Swedish Nobel prizewinner Svante Arrhenius in 1918 imagined the planet was covered in seas, swamps and steamy jungles perhaps populated by primeval monsters.

Astronomers of the 1930s and 1940s had different ideas. With the spectroscope they had discovered that the chief constituent of the atmosphere was the heavy gas carbon dioxide. This suggested that radiation from the Sun would be trapped by a "greenhouse" effect creating high temperatures which, some believed, might attain the boiling point of water. One theory was that Venus was in the throes of extensive vulcanism and that its thick atmosphere was volcanic dust suspended in a perpetually warm fog.

Venus: a hostile environment

Radar signals bounced off the planet from Earth in the 1960s indicated a rugged surface and possibly large craters. By finding a fixed point of reference these radar soundings showed that Venus turned on its axis in a reverse direction to Earth – once in 243 days.

Space probes – beginning with the fly-by of America's Mariner 2 in 1962 and continuing with the Soviet Venera capsules which entered the atmosphere – produced clear evidence of surface temperatures far above the melting-point of lead and atmospheric pressures that would crush any normal type of spacecraft.

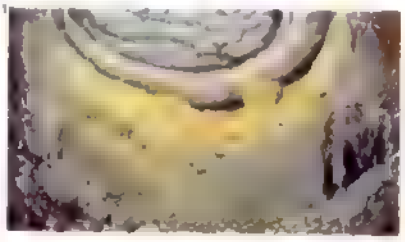
In 1975 Veneras 9 and 10 swung into orbit around Venus after releasing camera-equipped capsules which landed some 2,200km (1,370 miles) apart. Ruggedly built to withstand extremely high pressures they descended with the help of parachutes and a circular air brake. They were also specially cooled and insulated to work for at least half an hour on the fiercely hot surface. The first capsule – transmitting to Earth via its orbiting mother craft – worked for 53 minutes. It sent a panoramic picture of its surroundings showing a scattering of sharp-edged rocks 30 to 40cm (12 to 16in) across which appeared to be little affected by heat or wind erosion. The rocks seemed to be comparatively young and of a kind which may have been produced by subsidence or a meteorite impact. Instruments showed the pressure of the atmosphere to be 90 times that on Earth, the temperature 485 C (900 F). The second capsule, which transmitted for 65 minutes, showed a different landscape typical of old mountain formations. The rocks resembled huge pancakes with sections of cooled lava or the debris of weathered rocks in between. Here the pressure was 92 atmospheres, suggesting that the capsule had come down in the area of a valley.

The US Pioneer Venus Orbiter probe revealed the geography of Venus for the first time. After going into orbit in December 1978, its radar set probed the surface of Venus through the clouds, and showed up its relief. Venus consists largely of rolling plains

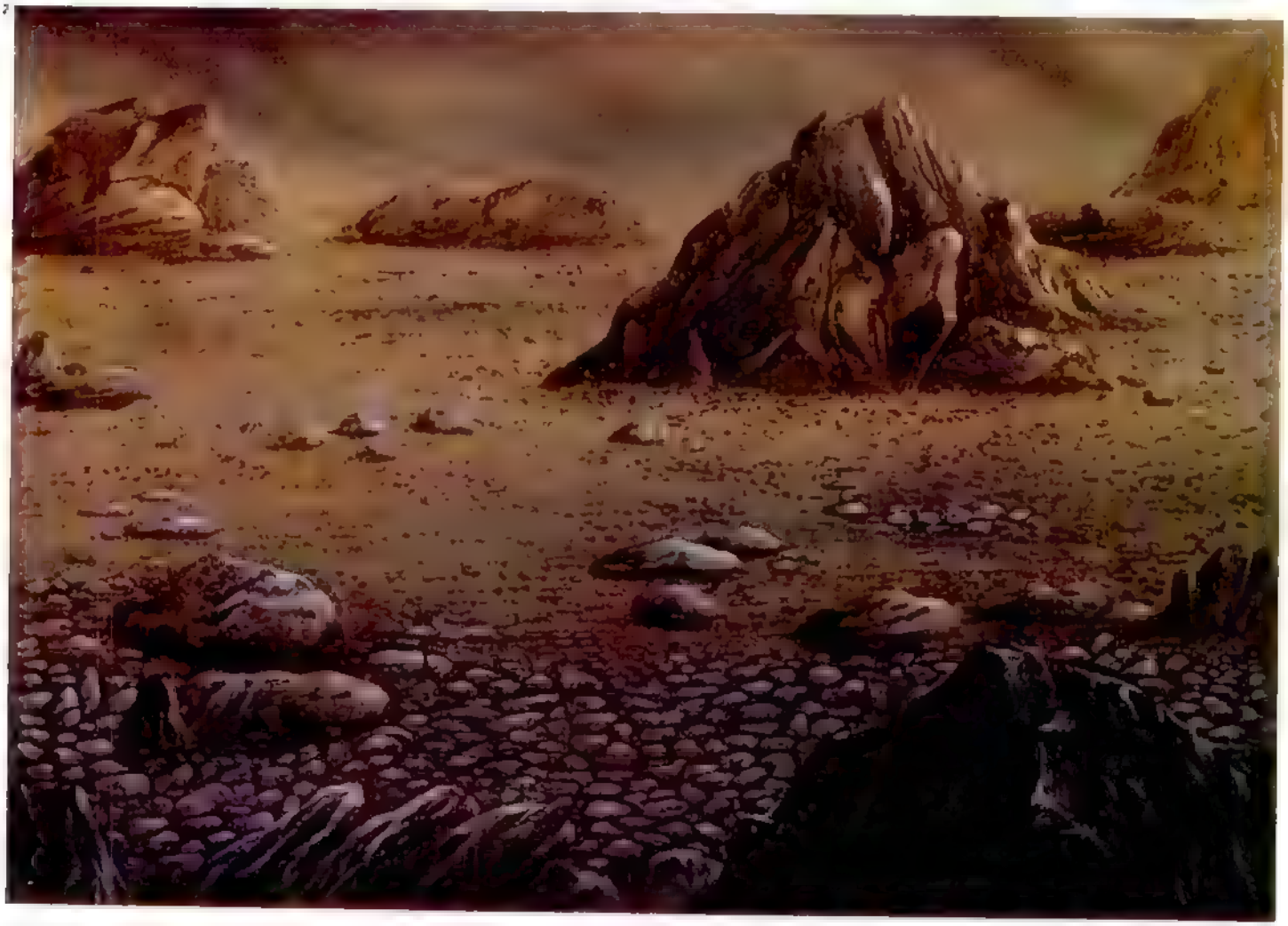
CONNECTIONS

See also

1 The dense atmosphere of Venus led scientists to believe that it must be super-refractive, that is to say the rays of light would be bent to such an extent that an observer on the planet's surface would have the impression of being in a vast bowl, with the horizon curving upwards all around him. However, the Venera 9 and 10 pictures showed that this is not the case. The atmosphere of Venus does not show the super-refraction that was forecast



2 Venus has proved to be a world quite different from anything that had been expected. The pictures from Veneras 9 and 10 show what the Russians have called a "stony desert", shown in this artist's impression. The rocks are relatively smooth, and it is thought that the erosion on Venus is less than on the Earth or even Mercury. The open sky can never be seen through the dense corrosive clouds of acidic vapour that surround the planet making Venus a gloomy, hostile world. As a potential colony, Venus has proved to be a disappointment. The surface features are almost certainly volcanic in origin. It is not known whether or not active vulcanism is in progress.



with two fairly small highland regions – one reaching a peak in Maxwell Montes 10,800m (35,400ft) high, a volcanic area with two 4,000m (13,000-foot) cones, and some large, deep valleys.

Both US and Soviet probes have contributed to our knowledge of Venus's cloud layers. The thickest clouds occur at a height of 50km (30 miles), although there are patchier regions of cloud and haze above and below. The clouds are composed mainly of sulphuric acid, with some yellowish sulphur crystals. Together with traces of water vapour in the atmosphere, the clouds and haze produce a greenhouse effect almost as strong as that due to the carbon dioxide constituting Venus's atmosphere. Winds of 360km/h (225 mph) whirl Venus's equatorial clouds around in only four days, sixty times faster than the planet's rotation.

Mercury: solar observation point

Mercury, the nearest planet to the Sun, takes nearly 88 days to make one revolution at an average distance of 58 million kilometres (36 million miles). A little less than half Earth's diameter, the planet turns slowly on its axis, completing a day in 58.5 Earth-days. Its surface is alternately baked by the Sun and frozen by the cold of outer space.

Mariner 10 – which flew past Mercury in 1974 after skirting Venus – discovered a Moon-like world with craters, mountains and valleys. There was a trace of atmosphere and no detectable magnetic field.

The asteroids: beacons in space

Mercury and Venus are the only planets moving at a distance from the Sun less than that of the Earth. Farther out, Mars is our nearest neighbour and then comes the main zone of asteroids or minor planets [Key]. The largest of these bodies, Ceres, is about 1,000-1,200km (600-750 miles) across, but most are so small that contacts with them would be more in the nature of docking operations than landings. Some asteroids wander far from the main belt. Icarus, for example, goes within 28 million kilometres (17 million miles) of the Sun, while Hidalgo swings out almost as far as Saturn. One day spacecraft from Earth may visit them.

KEY

The asteroids or minor planets may one day be the object of an exploratory mission. In this artist's impression the astronauts have docked with Eros and are setting up an inflatable, semi-transparent dome. They are preparing to make a geological survey of the asteroid. Eros is about 27km (17 miles) in diameter, and it is irregular in shape. Like most of the asteroids, so that its horizon slope seems strange. The orbit of Eros takes it away from the main belt of the minor planets and it may approach to within 24 million km (15 million miles) of Earth. Its surface is pitted with craters resulting from collisions with debris in the asteroid belt.

3

3 Mariner 10 approached Mercury in March 1974, seven weeks after by passing Venus, and sent back the first close-range pictures, showing that Mercury like the Moon, is cratered. In this picture Mariner 10's two "paddles" spanning nearly 9m (29.5ft) are seen: these provide the solar power for the various on-board systems. The dish antenna transmits information to Earth, and there are devices to measure magnetic fields, charged particles and ultra-violet and infra-red radiation. Mariner passed Mercury three times and on each occasion produced valuable data. It will continue to orbit the Sun indefinitely, although its "useful" life was over after it passed Mercury in the spring of 1975.

4

4 The planet Mercury has virtually no atmosphere and it is just as hostile as the Moon. In this picture the Sun is hidden behind a plug of lava from a volcanic eruption millions of years previously and since worn down by alternate expansion and contraction as a result of Mercury's great diurnal temperature range. Craters are seen on the surface, and the bright star in the upper right is the Earth-Moon pair. Mercurian days and nights are long because the planet rotates so slowly on its axis, and it is a desolate, lifeless world.

Exploring Jupiter and Saturn

Planetary scientists will never again live through a period as exciting as the two and a half years from early 1979 to 1981. In this period, the two US Voyager space probes passed Jupiter and Saturn. They not only laid bare the secrets of the two giants of the Solar System, but revealed for the first time more than twenty other worlds – the moons of these planets – known previously only as specks of light.

Some of these satellites are essentially planets – at least in size, even if they do not orbit the Sun in their own right. Jupiter's four Galilean satellites and Saturn's giant moon Titan all rival, or exceed Pluto in size. Jupiter's Ganymede is three-quarters as large as Mars, it ranks as the eighth largest world in the Solar System (excluding the Sun). Between them, the Voyagers trebled the number of explored celestial objects in the Solar System. An expedition like theirs, to Jupiter or Saturn, is rather like a trip to another planetary system, but with the considerable advantage that the different worlds are all placed conveniently close together.

Future unmanned space probes

The United States has firm plans to send two probes, code-named Galileo, to Jupiter in the late 1980s. One Galileo craft will go into orbit round Jupiter. For a period of almost two years, it will photograph the changing patterns in Jupiter's cloud layers and monitor the other worlds in Jupiter's family. The other probe will descend gently into Jupiter's clouds on a parachute, and relay information on conditions in the cold and windy hydrogen-helium atmosphere.

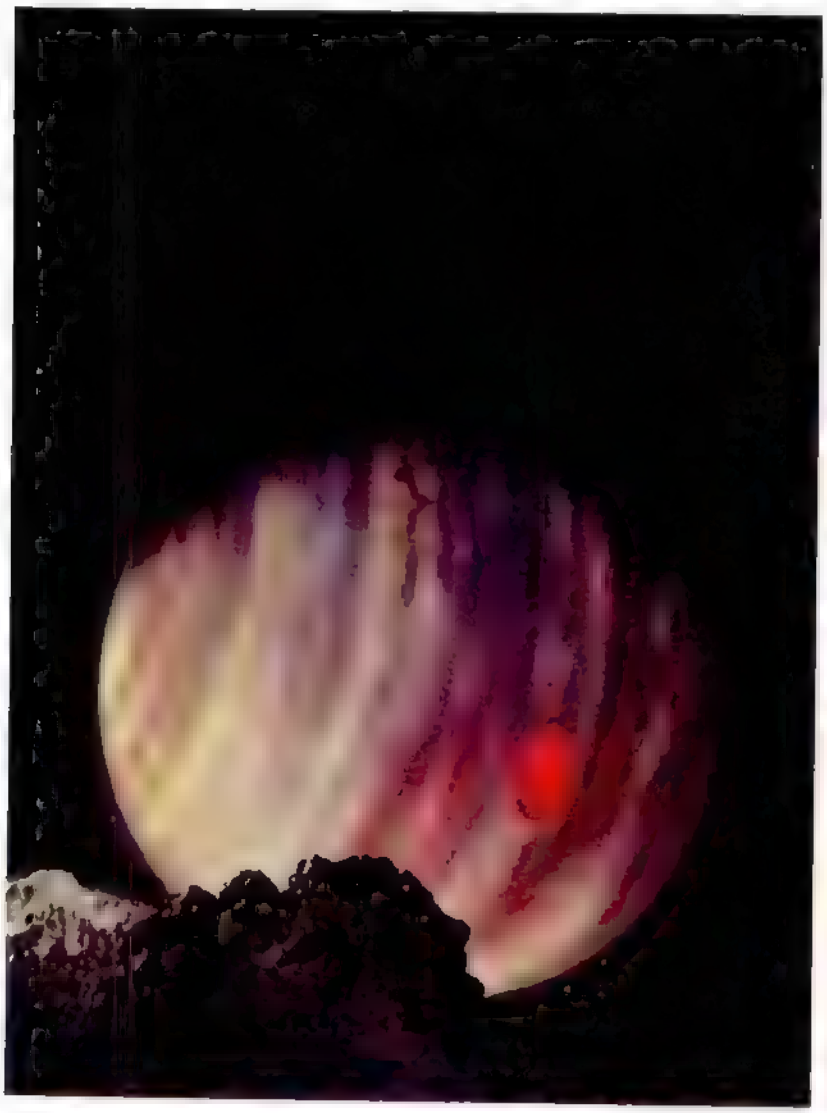
The Galileo probes may, however, still be cancelled because of the financial problems that are hitting the US space programme in the 1980s. Although the Space Shuttle will be available to launch them into Earth orbit in 1985 or 1986, the Galileo probes require a booster to propel them out to Jupiter – and funds for that rocket, a modified Centaur, may not be forthcoming. Even if they do fly, the Galileo craft may be the last probes to visit the outer Solar System in this generation, according to some financial analysts of NASA.

Funds permitting, there are exciting opportunities later this century. In the 1990s the planets are arranged to allow some easy and fast fly-by trajectories. Particularly interesting is a path that includes Jupiter, Uranus, Neptune and Pluto, arriving at the last planet in 2019. Alternatively, a spacecraft could use only Jupiter's gravity to send it straight to Pluto at higher speed, cutting the journey time to only eight years.

Scientists are however more interested in probes that will orbit planets, or descend into their atmospheres, rather than simply fly past. Galileo should achieve both aims at Jupiter. A similar space probe could reach Saturn in about five years. Once there, it could be swung into orbit by Titan's gravity. If properly aimed, the probe's orbit will continue to swing round, so it can keep making close passes to Titan and monitor its changing atmosphere too. After Voyager 1's exciting discovery that Titan has thick orange clouds of organic molecules, and possibly liquid methane seas, planetary scientists are now more keen to drop a probe into Titan's atmosphere than into Saturn's. This might be an amphibious lander, con

CONNECTIONS

See also



2 Jupiter, as seen from Amalthea would look magnificent with the shadows of Io and Europa clearly visible. Officially known as Satellite 5, Amalthea is the innermost member of the Jovian family, lying only about 110 000km (70 000 miles) from the planet's surface. It moves at a mean distance of 181 000km (112 000 miles) from Jupiter's centre. Amalthea is only 200km (124 miles) in diameter and is possibly distorted in shape by the intensely powerful pull of gravity from the planet. Expeditions to Amalthea could prove highly dangerous because it lies within the radiation zone of Jupiter.

1 Jupiter would dominate the sky to an observer standing on one of its four large satellites. He would clearly see the dark belts, the bright zones and the Great Red Spot. Io lies 422 000km (262 000 miles) from the centre of Jupiter, which is rather more than the distance of the Moon from the centre of the Earth. But the revolution period of Io is only one day 18.5 hours because Jupiter's powerful gravitational pull causes it to move much faster in its orbit. Io moves in the Jovian Magnetosphere and affects radio emissions from the planet. Parts of Io's surface may well be ice-coated.

Manned expeditions

[illegible] travel to Jupiter or Saturn, they [illegible] the Galilean satellites [illegible] surface [illegible] the base in the [illegible]

All the four Galilean satellites of Jupiter [illegible] would [illegible] more hi mely than the free fall of [illegible] power out in this frozen [illegible] the Solar System [illegible] but only if the lo- [illegible] will not [illegible]

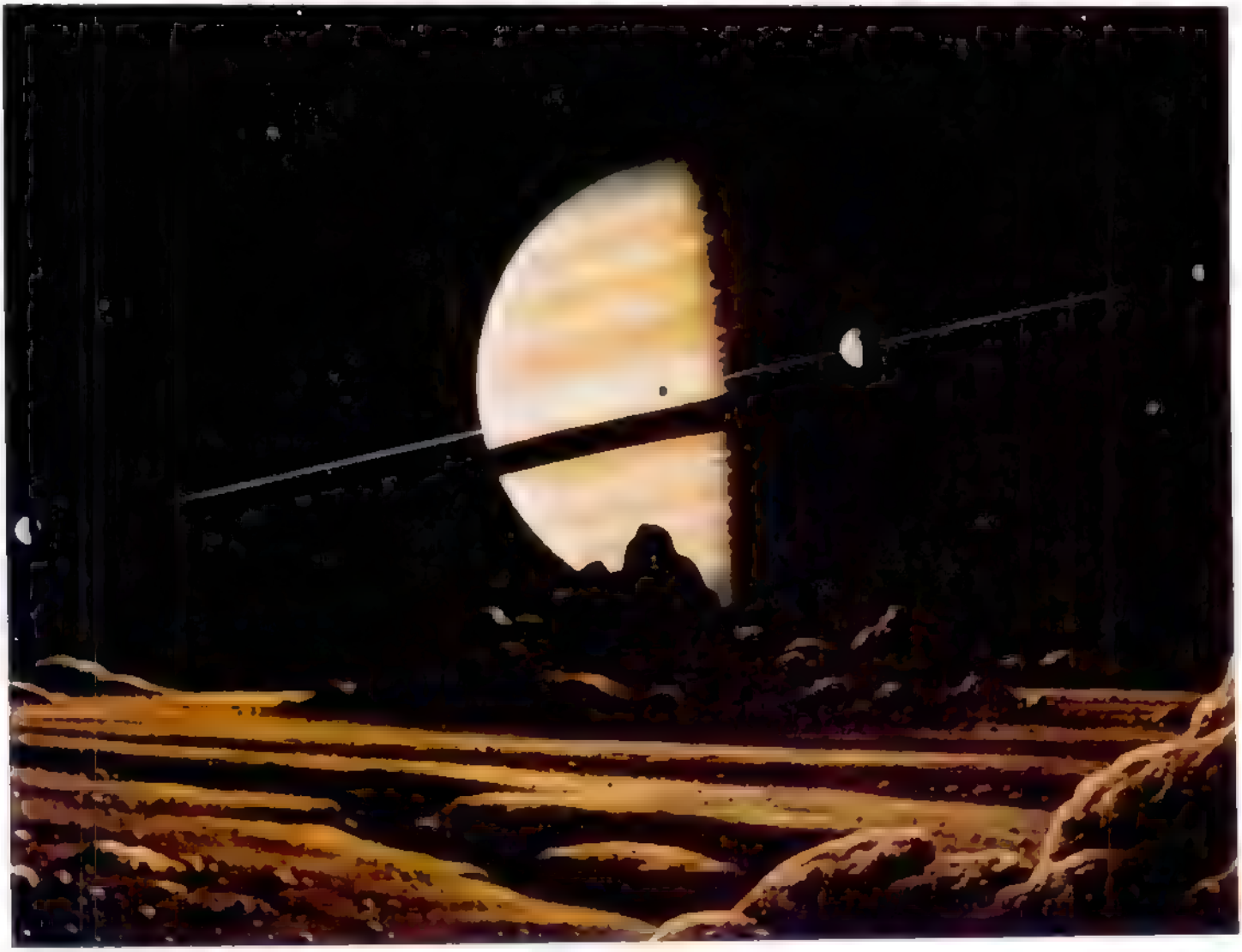
The next in moon, small Amalthea, [illegible] ideal base for observing Jupiter itself [2]. Unfortunately, Amalthea [illegible] Galileans all lie within Jupiter's [illegible] radiation belt, which would [illegible] without proper shielding.

Saturn offers a wider range of moons for [illegible] ionization, and a much less intense [illegible] Titan, largest moon R[illegible] 3 [illegible] magnificent views of Sa[illegible] twin in size, Iapetus, which lies outside Sa[illegible] together.

The alternative would be a base on T[illegible] unbroken pressing range clouds. Although Titan's atmosphere is unbreathable and b[illegible] more than the Earth's, so a manned base would not need the strong h[illegible] base built in the near[illegible] moons surfaces. In addition, T[illegible] sphere would shield the base from rad[illegible]

Certainly, Titan will be a weird world to explore. Its solid surface "rocks" are made [illegible] water ice, and they are lapped by oceans of liquid methane. Near the [illegible] glaciers of frozen methane cut through [illegible] surface. Some scientists believe that organic droplets must rain down from the orange clouds to [illegible] Titan [illegible] settlers will use this as a basis for [illegible]

Saturn surrounded by a host of stars appears to the [illegible] [illegible] bright star. Although [illegible] remote, it is closer than any star and lies within probe range.



3 Saturn, seen from Rhea, would display [illegible] satellites. [illegible] yellow light would cast a bright glow [illegible] Rhea's surface. Rhea's [illegible] sixth largest satellite, from Saturn [illegible] the planet at a distance of [illegible] 527 000 km – 327 [illegible] miles) from its centre [illegible] Rhea's Saturn's [illegible] are 8 [illegible] 500 km [illegible] miles. Rhea takes four days [illegible] 12.5 hours to [illegible] perfect [illegible] the Moon [illegible] face except that it [illegible] rocks [illegible] from the two outermost members [illegible] Rhea [illegible] Saturn [illegible] Which [illegible] fore, from Rhea, the [illegible] rings would appear as a thin [illegible] light [illegible] away [illegible] edge [illegible] A [illegible] server would never be able to see the Cassini Division [illegible]

Exploring the outer planets

In January 1986 the US space probe Voyager 2 will pass the distant planet Uranus. It will be its victory – the probe did have problems with a jammed camera platform as it passed Saturn in 1981 – Voyager 2 should return up-to-date views of an outer giant planet. When it arrives, the Voyager will have been on the wing for 8½ years – and if this seems like a long enough time, the journey would have taken far longer if Voyager 2 had not been launched on one of those rare occasions when it could use the gravity of both Jupiter and Saturn to speed it on its way. Uranus is such a distant world that it is difficult to imagine that astronauts would ever travel there by the same kind of technology that propelled the Voyagers.

It is much too early to speculate about when the first manned expedition will venture as far as Uranus. Vehicles much more sophisticated than those planned at the moment will be needed, even so landing on the planet's surface would be impossible. Uranus, like Jupiter and Saturn, has a surface of gas, although the composition of the planet differs in various important details. There is no evidence so far to indicate the existence of dangerous radiation zones like those of Jupiter. Of the five satellites, Miranda is the closest to the planet (only 130 000km – 80 000 miles – off) but it is likely that the first landing will be made on one of the larger satellites, such as Ariel [1]. Because of its strange axial tilt, Uranus sometimes appears as a crescent with its horns extending from one side of the equator to the other rather than from pole to pole – a case unique in the Solar System. The strange greenish light of Uranus upon the rocks of one of its satellites will make an eerie picture and one of emptiness and desolation.

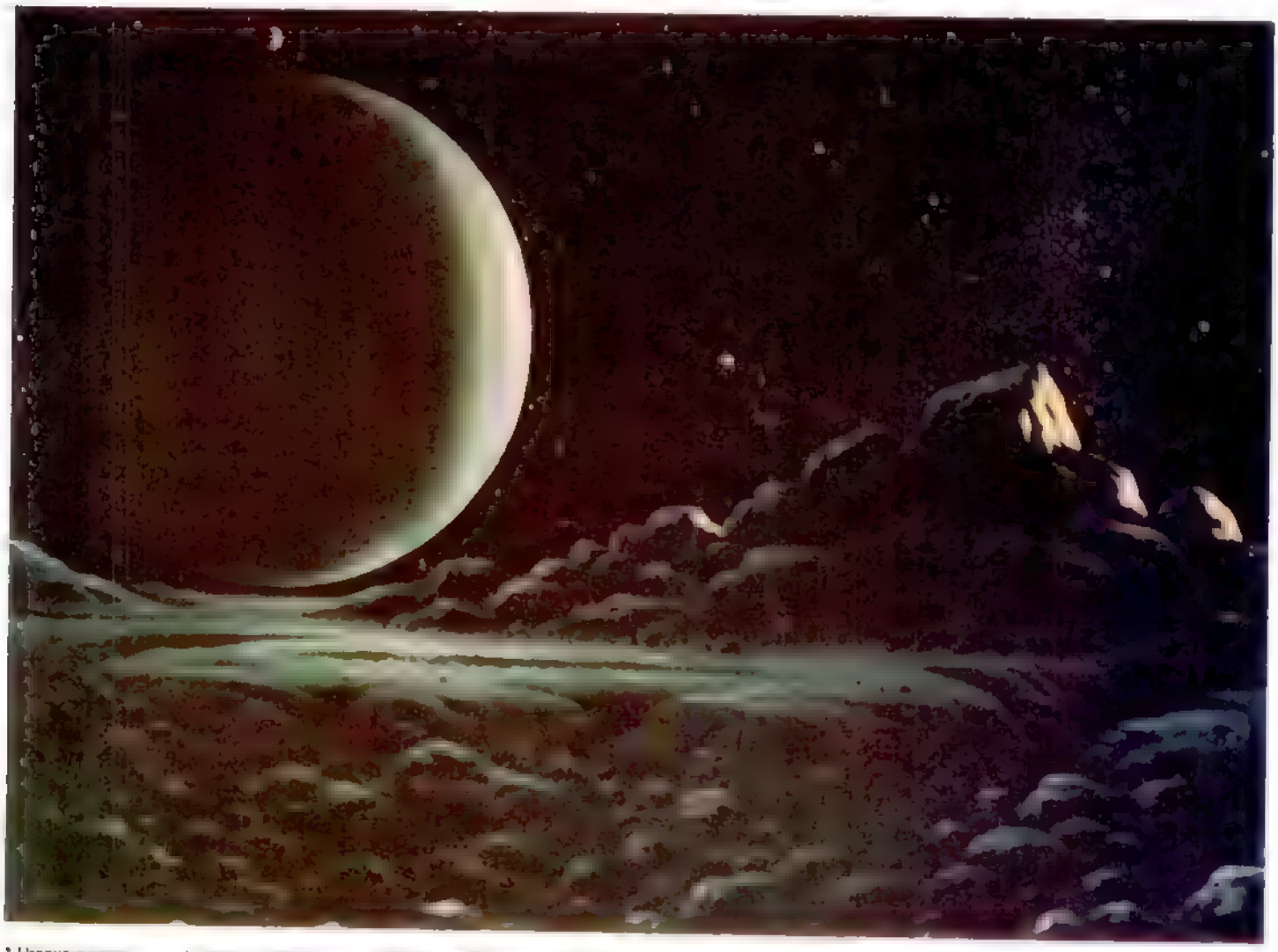
Neptune and Triton

Beyond Uranus lies Neptune – but again the distances involved are immense, the distance of Neptune from Earth is 5 times that of Uranus [2]. But at least there is a more promising satellite, Triton, which is much larger than any of the attendants of Uranus and may possibly have the same kind of atmosphere as Titan, Saturn's largest satellite, although no proof of this has yet been obtained. Triton is unique in one large satellite in that it moves round its primary [illegible] Neptune [illegible] Because [illegible] Neptune [illegible] Triton [illegible] Neptune.

If a major outpost is to be established in these desolate regions of the Solar System it is most likely to be on Triton. The other satellite of Neptune, Nereid, is extremely small – less than 300km (200 miles) in diameter – and with its highly eccentric orbit would offer few advantages as an observation base. Even from Triton the other planets would not be seen to advantage: only Uranus would seem brighter than it does from Earth and it would be an inner planet keeping in the same area of the sky as the shrunken but still intensely bright Sun. Moreover, when Uranus and Neptune lie on opposite sides of the Sun an observer on Triton would find

CONNECTIONS

See also



1 Uranus is seen here from its satellite Ariel in this artist's impression. A probe leaving Uranus from Earth would be launched to the neighbourhood of Jupiter that powerful [illegible] gravity would then be used to pull the probe and a [illegible] [illegible] Jupiter and [illegible] towards Uranus. The gaseous surface of Uranus [illegible] prohibits any landing. But landings may be possible on some of its five satellites. Of these, the closest to Uranus, Miranda, is very small. The largest of the main satellites is Ariel, which moves round Uranus at a distance of 192 000km (119 000 miles) from the centre of the planet in a period of 2 days 12 hours 29 minutes. Nothing is known about the surface of Ariel, but its size seems to be considerably smaller than the Moon – about [illegible] in diameter. Ariel and the other satellites moves in the plane of Uranus's equator. In this view Uranus looks like a crescent but the horns extend from one side of the equator to the other. With its remarkable inclination of 98° in 1985 the pole of Uranus will face the Sun so that there will be a polar day lasting 42 earth years. As from Ariel or any other satellite Uranus will appear as a half disc. The changing surface details will be displayed although the pale greenish disc of Uranus is markedly less active than either Jupiter or Saturn and less interesting visually.

Uranus difficult to see over a period of years. Saturn would be even more elusive because the distance between Neptune and Saturn is much greater than that between Saturn and Earth and the inner planets would be virtually out of view. Yet from Triton there would be no interference from brilliant solar radiation and a base there could still make useful observations outside the Solar System and thus contribute substantially to man's knowledge of the universe.

The outermost planet

Little is known about Pluto, the outermost planet [3]. It is slightly smaller than Triton and has a surface of methane ice, which tends to confirm current theories explaining the origin of the Solar System and the sequence of planetary formation. At perihelion, or closest point to the Sun, it comes within the orbit of Neptune, the next perihelion passage is due in 1989. At aphelion, Pluto is more than 7 000 million kilometres (4 500 million miles) from the Sun.

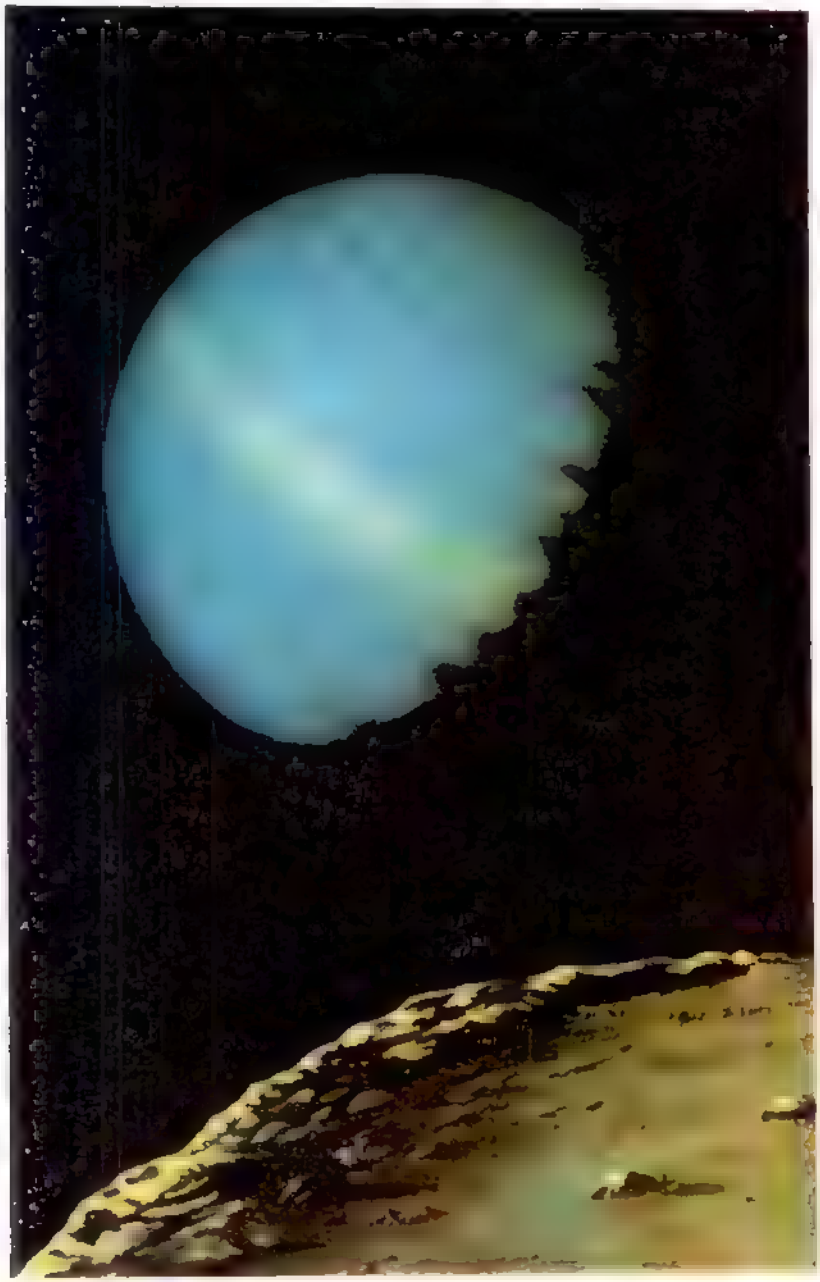
A future spacecraft, even an unmanned probe, will also send back information on Pluto's moon, Charon. Any clues to the formation of this double-planet system will help with understanding the origin of the Earth and Moon. If astronauts ever land on Pluto, they will find that the Sun looks no larger than Jupiter does as seen from Earth, although it will still shed a certain amount of light over the bleak Plutonian surface.

Communications with Earth will be slow. A radio wave would take about five hours to cross the space from Pluto to Earth, so that if a message is transmitted from Earth, a reply would take ten hours.

Exploring the comets

Although Pluto is the outermost planet, there may well be opportunities for studying material from even farther out in the Solar System. Comets [Key], those wraith like and insubstantial objects, mostly have very eccentric orbits and there is a real possibility of sending a probe through a comet that has come from the region beyond the orbits of Neptune and Pluto. At the next return of Halley's comet in 1986 several probes will investigate it at close quarters.

Comets, the erratic wanderers of the Solar System, can journey farther than the known outermost planet, Pluto. Shown here is a probe approaching a comet on its journey; the Earth and Moon appear in the upper right. Unlike a planet a comet is not a solid, massive body; it is composed of relatively small particles, mainly icy in nature, together with extremely tenuous gas. Hence there is no reason why a probe should not be able to pass right through it. The tail is particularly tenuous, thus the background stars can be seen through it virtually undimmed. Some comets are known to travel a distance equivalent to a third as much again of Pluto's distance from the Sun.



2 Neptune, unlike Uranus, has a normal axial inclination of 29°, less than 6° greater than that of the Earth. Surface detail is on its bluish disc and difficult to make out, but there seems little doubt that Neptune and Uranus are similar. The illustration shows a view as it would appear from Nereid, the smaller of Neptune's two satellites, which has an eccentric orbit; at its closest to the planet it approaches to within 1 400 000km (870 000 miles) and this is depicted in the illustration. Nereid takes almost one Earth year to complete a full revolution around the planet Neptune.

3 Pluto at times comes within Neptune's path because of its relatively eccentric orbit. The next perihelion is due in 1989; for some years on either side of that date Pluto will no longer be the outermost known planet. At aphelion it recedes to more than 7 300 million km (4 600 million miles) from the Sun. The temperature of Pluto is estimated to be as low as −230 C (−382 F). So far no atmosphere has been detected; it possesses a surface of methane ice; its composition may well resemble that of the moons of the outer planets rather than that of the planets themselves.

Beyond the Sun's family

Exploration of the Solar System is progressing and, provided that the present rate of progress is maintained, all the planets will have been contacted by automatic probes within the next 50 years — probably well before — and manned expeditions will have been sent to those worlds like Mars that are not overwhelmingly hostile and are within reasonable range. Yet even when mankind has finished his exploration of the Solar System, he will hardly have scanty space the vast bosom of the universe.

See also

Problems of interstellar travel

The Solar System is only a small part of the universe. If the distance between the Earth and the Sun were represented by 2 cm, then the nearest star would be almost 7km (4.3 miles) away. Stellar distances are so great that man's present technology cannot yet begin to bridge them. Although two interstellar probes have been despatched — Pioneer 10, which by-passed Jupiter in December 1973, and Pioneer 11, which did so about a year later — neither will approach a star for many thousands of years. Neither can transmit signals to Earth. They are probes launched into space in the hope that they may reach a world of advanced beings who can communicate with Earth — and in the hope that when that happens mankind will have advanced sufficiently to be able to understand that communication.

Even if travelled at the speed of light a probe would take more than four years to reach Proxima Centauri — the nearest star that is sufficiently like the Sun to have a family of planets moving round it. According to the theory of relativity, which has so far survived every test, it is impossible for any material body to travel at the speed of light and any spacecraft that can be planned in the present state of technology is by comparison very slow indeed.

Journeys to other solar systems will have to be made by methods that are as yet unknown and will probably need technologically advanced assistance. Wonderful ideas have been suggested in the days of Jules Verne. All kinds of suggestions have been made. A favourite scheme, for instance, is that of the spacecraft in which those who set out die at an early stage of the voyage, leaving their descendants to finish the journey. Alternatively, it has been suggested that the travelers should be put into a state of suspended animation, being conveniently woken up just before arrival at a suitable planet. Then there are theories involving telepathy and teleportation — transporting matter through space in much the same way as television transmits pictures. They are all intriguing but present beyond man's powers to implement.

All that can really be said is that interstellar travel is impossible by any known method. However, a spectacular breakthrough may come eventually, and it is conceivable that alien beings may visit the Earth before man has developed sufficiently to be able to visit them.

Finding new planets

What are the prospects of finding planets orbiting other stars? The Sun is a normal G-type dwarf and there is no reason to regard it as exceptional in any way. Moreover G-type dwarfs are common and it thus seems that planetary systems are likely to be abundant in

1 The Moon can be reached from Earth in a few days and to send a rocket to Mars or Venus involves a journey of only a few months. But interstellar travel presents a totally different picture and possibly many problems. The distances involved are millions of times greater than those used today would be necessary even for flights to any star. In the USA research on such has been done on the possibility of building what is termed a photon rocket, in which the gases emitted from the exhaust are replaced by a stream of photons that is, in effect, a beam of light. The thrust produced would be low, but it could be maintained indefinitely and over a period of years the acceleration would build up until the velocity of the vehicle approached that of light. According to relativity theory the proposed velocity of light can never be reached by a material body, since it would involve infinite mass. In principle the photon rocket can be compared to a gigantic electric torch which is driven forward by the light it emitted. One possible design is shown here. The rocket would be more than 9.5km (6 miles) long with a crew of 300–500. Our Galaxy is in the background in the inset, lower right. Our Galaxy is shown first face on and then edge on, with the position of the Solar System indicated by a red circle. Even at the velocity of light it would take a probe 100 000 years to pass from one side of the Galaxy to the other.

the Galaxy. Other kinds of stars are less promising. For instance, a red giant star that has left the main sequence and has swelled to many times its original size is likely to have swallowed up any planets it may once have had [2], while a hot, massive blue or white star will have run through the earlier part of its evolution so quickly that planets that resemble the Earth will hardly have had time to develop.

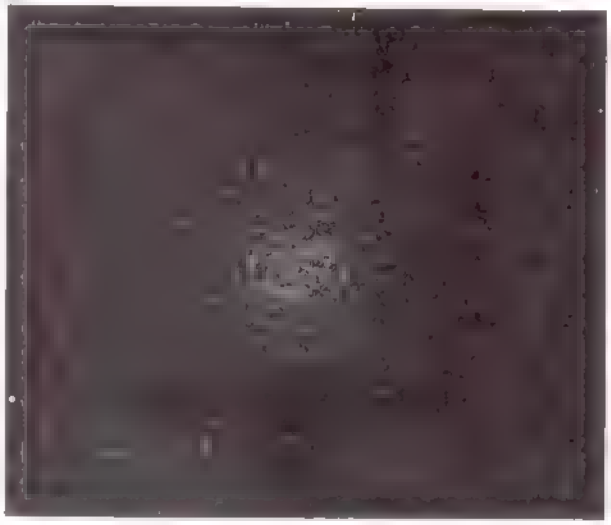
Then there are the faint red stars, so feeble that they have never joined the main sequence and are on their way to extinction. Barnard's Star, at a distance of only a little more than five light-years from Earth, is one of these. As it moves through space it "wobbles" slightly and there have been suggestions that it is being pulled out of position by an orbiting planet or planets. If so, what would such a planet be like?

An alien environment

By Earth standards, a planet associated with a star like Barnard's would be a dreary world. Its only light would come from a dim red sun and it would therefore be cold. Any life there would have to contend with an environment that man would find intolerable. Yet it would be unwise to dismiss such planets as possible places for colonization.

Barnard's Star is not active by stellar standards. Stars that are much more violent become unstable as they use up their nuclear energy and some produce nova-like outbursts [3]. The effects of these upon any orbiting planets would be catastrophic. If our Sun became a nova, all life on Earth would be destroyed in a matter of hours. Luckily this cannot happen yet, because the Sun is a stable star and will not significantly change its structure for at least 5,000 million years. There must, however, have been planetary systems that were engulfed and destroyed by their dying central stars.

There would have been a period of warning in which the inhabitants of a threatened planet, if they were sufficiently advanced technologically, could take certain steps to save themselves. The obvious step would be into space – abandonment of the world about to die in favour of a new start on a new world, whatever the difficulties.



In this cluster of stars NGC 6819 in Cygnus, photographed with the 200-in (508-cm) reflector at Palomar, each star is a sun in its own right. Many may have associated planet families, but there is no direct optical proof, no telescope yet built or planned can hope to show a planet of another star.

2 Zeta Aurigae is a binary system made up of a vast red supergiant together with a much smaller white star about to be eclipsed. The red supergiant is an old star that has left the main sequence and has swelled out to a diameter greater than that of the orbit of the Earth around the Sun. Because it has become luminous it will have raised the temperatures of its planets to intolerable levels and any inner planets of its original system will have been destroyed. No life can be expected in a system of this kind. The illustration shows a view from a hypothetical planet.

3 This barren, dead planet is on the outskirts of a hypothetical planetary system – the "Pluto" of the system. The central star, a binary system, has flared up as a nova, resulting in a tremendous, although temporary, increase in luminosity. The planet's surface has been scorched by the tremendous radiation; its water has evaporated and even its atmosphere has been driven away so that no life can possibly survive. The inner planets of the system have been completely destroyed. The sky is enriched by an aurora-like display caused by a shell of gases released by the star as its outer layers expand. When the nova outburst is over, the planet will remain cold and sterile, clinging the feeble remnant of its once glorious sun.

Worlds of many kinds

Life on Earth has developed along its familiar lines because the conditions are suitable for it. If the Earth were smaller, colder or less massive, then life would have taken on different forms, and if the conditions were unsuitable, no living organisms would have developed at all. Life, wherever it is found, is suited to its environment. If a star like the Sun were attended by a planet the size and mass of the Earth orbiting it at a distance of 150 million kilometres (93 million miles), then Earth-type life might reasonably be expected. In 1972 Pioneer 10 was launched to probe beyond the Solar System. It carried a plaque [Key] to communicate with any intelligent life that might encounter it.

Alien life forms?

This does not mean that all forms of life must correspond with the terrestrial pattern. There is nothing, in theory, against an intelligent astronomer having six legs and two heads. If he were made up of the same materials as ourselves he would not, despite his appearance, qualify as one of those interesting creatures novelists call BEMs, or "bug-eyed monsters". This term is reserved for entirely alien creatures, breathing pure methane, say, and able to survive in temperatures of −150°C or below. It cannot be said definitely that alien life does not exist, all that can really be done is to take the available facts and then put the most reasonable interpretation upon them. When this is done the existence of bug-eyed monsters begins to appear improbable.

Rational discussion of life on other planets must be confined to "life as we know it". Any extension to include alien life forms means that speculation becomes not only endless, but also pointless.

An Earth-type planet may be expected to produce Earth-type life, essentially similar to our own and no doubt subject to the same weaknesses. For instance, the star Delta Pavonis, at a distance of 19 light-years from the Earth, is strikingly similar to the Sun: we have no idea whether or not it has a planetary system, but there seems no reason why it should not be attended by a world similar to the Earth, in which case its inhabitants may at this very moment be speculating about the possibility of intelligence upon a planet orbiting a fourth-magnitude yellow star in their own sky. If this hypothetical planet lay farther from Delta Pavonis than the Earth does from the Sun, the colder climate would produce life forms more akin to those in our polar zones; if it lay closer, life would be more equatorial. Of course, this can be no more than speculation because nobody knows whether a planet capable of supporting life will actually produce it, but again there seems to be no valid reason why it should not.

Different perspectives

Many stars in the Galaxy are members of binary systems and it is fascinating to picture a planet lit by two suns – perhaps of different colours, one yellow and one blue, giving strange, spectacular colour effects. Then there are the variable stars, some of which are completely regular while others are violently explosive. It is hard to see how a variable star could be attended by a life-bearing planet because there would be extreme fluctuations in climate, but most variable stars are well advanced in their evolution, so that life on

CONNECTIONS

See also

1 A planet of Proxima Centauri, the nearest star to the Earth, may exist, according to recent investigations. This picture of its surface is thus founded on something more than sheer imagination. Proxima Centauri is a dim red dwarf that has never joined the main sequence nor passed through the giant stage. It does not cast as much light as the Sun on its attendants. The planet thought to be moving round it is assumed to have an orbital period of between 10 and 12 days. The outer edge, or limb, of Proxima is not sharp, like that of our Sun, but diffuse, because the density of the outer layers is low. It is also assumed that the planet has a tenuous atmosphere. Because Proxima is a flare star, the climate of an inner planet will be unstable and the landscape extremely desolate. No life can be expected there. But water, shown in the picture as a lake fringed by glittering crystals of ice, may survive. The black circle represents a possible satellite as it would be seen in silhouette against the red disc of Proxima. In the sky is seen a familiar constellation pattern, the W of Cassiopeia, all of whose stars are remote and will therefore look much the same from Proxima as from the Earth. To the left of the W is another star – our own Sun – which will be easily visible with the naked eye.

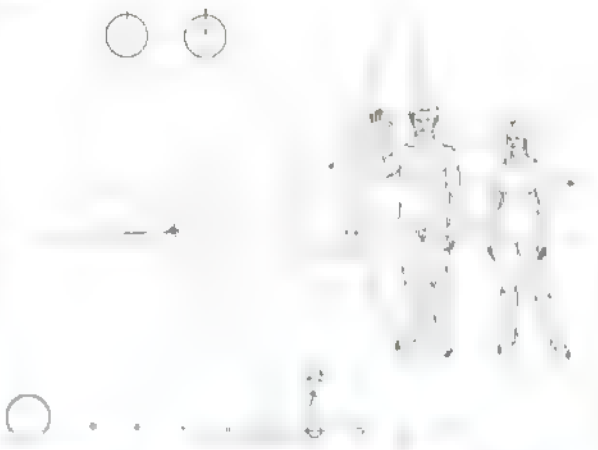
any of their surviving planets may have died out in the remote past

The Sun lies in a relatively sparsely populated region of the Galaxy, at the edge of one of the spiral arms (from which it emerged perhaps 5,000 million years ago). If the Sun lay in one of the rich globular clusters the sky would be ablaze; there would be many stars shining more brilliantly than Venus does to us and since globular clusters are "old" by stellar standards there would be many red stars which would have left the main sequence. There is no reason why stars inside globular clusters should not be attended by planetary systems. Astronomers there would be at something of a disadvantage, because it would be difficult for them to see clearly beyond the confines of their own dense cluster and they could know little about the greater universe beyond.

Interstellar communication

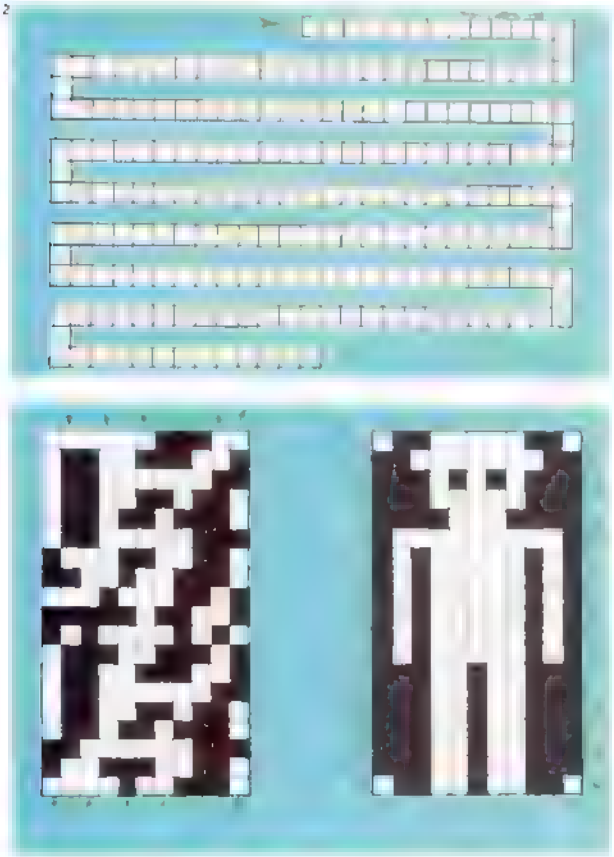
The only means of achieving communication with other planetary systems, using existing techniques, is by radio. Radio waves move at the same velocity as light and would take years to reach even the nearest star; nevertheless, the reception of a signal pattern rhythmical enough to be classed as artificial would be of unparalleled significance. Attempts have already been made to distinguish such a pattern.

In 1960 radio astronomers at Green Bank, West Virginia, began an ambitious programme known officially as Project Ozma. With powerful equipment they concentrated upon the two nearest stars that are reasonably like the Sun – Tau Ceti and Epsilon Eridani, both of which are slightly smaller and cooler than the Sun and are more than ten light-years away. A wavelength of 21 cm (8in) was selected because this is the wavelength of the radio signals emitted by the clouds of cold hydrogen spread throughout the Galaxy. It is logical to believe that other astronomers, wherever they may be, are also devoting their attention to this particular wavelength. The earliest that a signal from an alien world could be received by astronomers on Earth is 1980, but it may, and probably will, be many years after that before contact is made.



This plaque is carried on Pioneer 10 (launched in 1972) it will be the first vehicle to leave the Solar System. The radiating lines represent 14 pulsars, the binary notation gives the r frequencies relative to the universal constant (the hydrogen atom, upper left). The regular decrease in the pulsars' frequencies will give the time elapsed since launch. The Earth's position in the Solar System with Pioneer's path (bottom), and the male and female in proportion to Pioneer are shown.

2 Communication with alien civilizations could be achieved by means of mathematics – a system man discovered rather than invented. One suggested method is the transmission of two signals of different kinds (say dots and dashes). 209 signals are sent as 0's and 1's. A listener could represent the 0's by black squares and the 1's by white squares (or vice versa). 209 has only two factors: 11 and 19. Accordingly, a receiver has the choice of dividing the signals into 11 groups of 19 or 19 groups of 11. The second alternative gives the correct intelligible picture – that of a biped, as shown.



3 A radio telescope of the kind pictured here may well exist on a planet far away in the Galaxy and could be used to transmit codes in mathematical form. The parent star (bottom left) is like the Sun, and the planet itself is like the Earth, which means any forms of life found there might be similar to those now on Earth. A large nearby satellite is shown. The radio telescope is constructed along essentially the same lines as those already set up on Earth and could establish contact with an instrument such as the Haystack radio telescope at the Massachusetts Institute of Technology.

Astronomical tables and space milestones

THE SOLAR SYSTEM

Object	Distance from Sun in millions of km	[millions of miles]	Diameter in km	[miles]	Mass in Earth masses	Rotation period	Orbital period	Number of satellites
Mercury	58	36	4 880	3 032	0.05	58.7 days	88 days	0
Venus	108.9	67.20	12 100	7 500	0.82	243 days	224.70 days	0
Earth	149.59	92.950	12 756	7 926	1.00	23 hr 56 min 4 s	365.25 days	1
[illegible]	227.94	141.321	6 790	4 216	0.11	24 hr 37 min 23 s	686.96 days	2
[illegible]	778.38	483.6	142 800	88 700	318.9	9 hr 55 min	11.86 years	16 or 17
[illegible]	427	887	120 000	74 600	95.2	10 hr 40 min	29.46 years	17
[illegible]	2 869	1 780	52 300	32 400	14.6	approx 16 hr	84.0 years	5
[illegible]	4 496	2 793	49 500	30 700	17.2	approx 18 hr	164.8 years	2
[illegible]	5 940	3 690	3 500	2 200	0.002	6.39 days	250.3 years	1
[illegible]	384 000 km from Earth		3 477	2 160	0.012	27.30 days		
	238 000 miles							
Sun			1 392 000	865 000	333 000	25.38 days		9 planets

1 earth mass = 5.90 x 10^21 kg
1 [illegible]

BRIGHTEST STARS

Name	Apparent magnitude	Absolute magnitude	Distance [light-years]
[illegible]	-1.47	+0.7	8.7
[illegible]	0.71	5.5	300
Alpha Centauri	0.27	+4.6	4.3
Arcturus	0.06	0.3	36
Vega	+0.03	+0.3	26
Capella	0.08	+0.1	45
Rigel	0.15	-8.2	850
Procyon	0.34	+2.8	11
Achernar	0.49	-1.3	75
Betelgeuse	variable	variable	650
Hadar	0.61	-4.3	300
Altair	0.75	+2.2	16

NEAREST STARS

Name	Apparent magnitude	Absolute magnitude	Distance [light-years]
Proxima Centauri	10.7	15.1	4.3
Alpha Centauri	0.0	4.4	4.3
Barnard's Star	9.5	13.2	6.0
Wolf 359	13.5	16.5	7.6
Lalande 21185	7.5	10.5	8.2
Luyten 726-8	12.4	15.4	8.7
Sirius	-1.5	1.4	8.7
Ross 154	10.6	13.3	9.3
Ross 248	12.2	14.7	10.3
Epsilon Eridani	3.7	6.1	10.8
Ross 128	11.1	13.5	11.1
Lalande 789-6	12.2	14.6	[illegible]

ASTRONOMICAL DISTANCES

Name	Nature	Distance [light-years]
Vega	Star	26
Aldebaran	Star	65
Pleiades	Star cluster	410
Rigel	Star	850
Deneb	Star	1 500
M71	Globular cluster	6 000
Large Magellanic Cloud	Galaxy	180 000
Small Magellanic Cloud	Galaxy	200 000
Andromeda Spiral	Galaxy	2 200 000
M87	Galaxy	40 000 000
3C 273	Quasar	1 500 000 000
OQ 172	Quasar	10 000 000 000

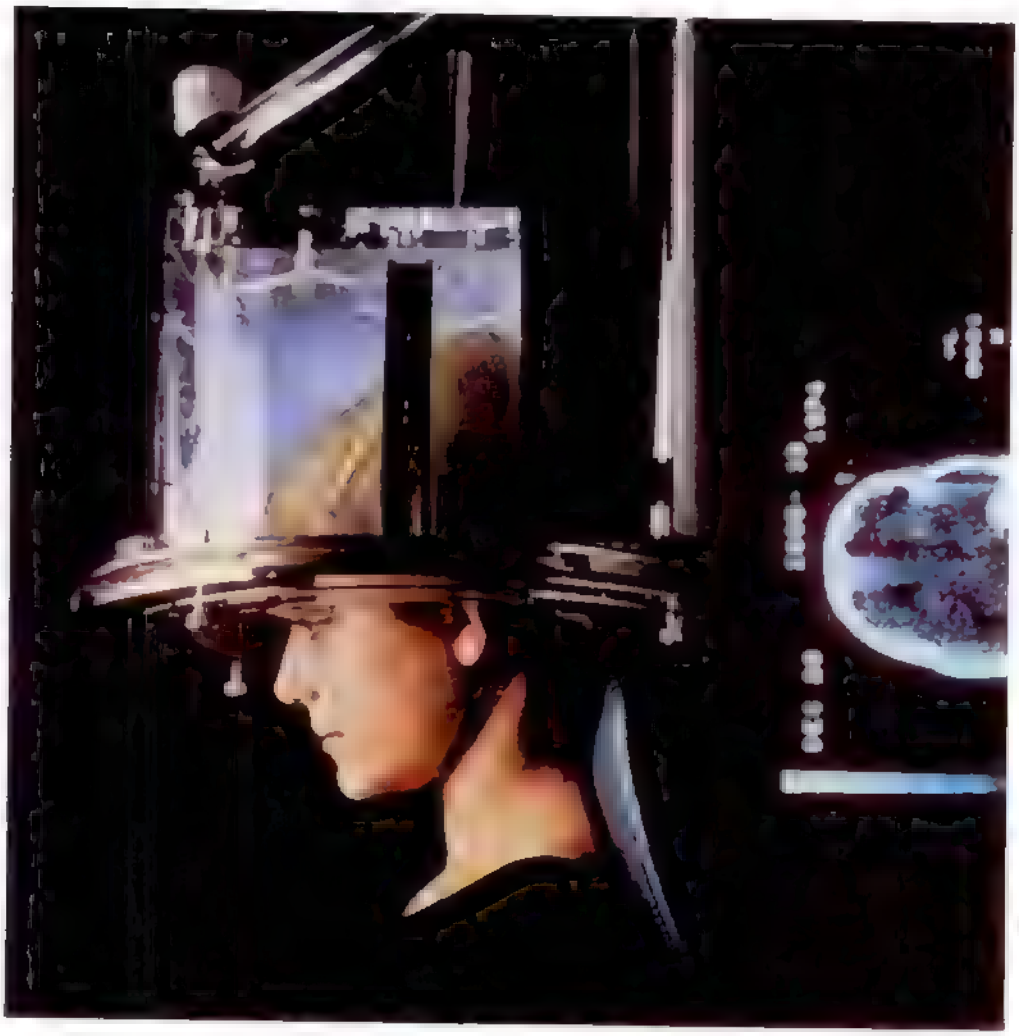
ASTRONOMERS ROYAL

Name	Dates
John Flamsteed	1675-1719
Edmund Halley	1720-1742
James Bradley	1742-1762
Nathaniel Bliss	1762-1764
Nevil Maskelyne	1765-1811
John Pond	1811-1835
Sir George Biddell Airy	1835-1881
William Henry Mahoney Christie	1881-1910
Sir Frank Watson Dyson	1910-1933
Harold Spencer Jones	1933-1955
Sir Richard van der Riet Woolley	1955-1971
Sir Martin Ryle	1972

MILESTONES IN SPACEFLIGHT

Spacecraft	Launched		Achievements
USSR Sputnik 1	4 October 1957	Unmanned	1st artificial satellite
USSR Sputnik 2	3 November 1957	Unmanned	1st inhabited spacecraft (the dog Laika)
US Explorer 1	31 January 1958	Unmanned	1st US satellite, discovered inner Van Allen belt
USSR Luna 1	2 January 1959	Unmanned	1st spacecraft to escape the Earth's gravitational pull
USSR Luna 2	12 September 1959	Unmanned	1st landing on the Moon
USSR Luna 3	4 October 1959	Unmanned	1st orbit of the Moon; 1st photographs of the far side
US Tiros 1	1 April 1960	Unmanned	1st weather satellite
USSR Vostock 1	12 April 1961	Manned	1st manned space flight (Yuri A. Gagarin)
US Mercury Redstone 3	5 May 1961	Manned	1st US manned space flight (Alan B. Shepard)
US Mercury Atlas 6	20 February 1962	Manned	1st US manned orbit of the earth (John H. Glenn)
US Telstar	10 July 1962	Unmanned	1st transatlantic relay of television transmissions
US Mariner 2	27 August 1962	Unmanned	1st fly-by of Venus; measured atmospheric and surface temperatures
USSR Vostok	16 June 1963	Manned	Valentina Tereshkova becomes 1st woman in space
US Mariner 4	28 November 1964	Unmanned	Mars fly-by; 1st photographed surface and studied the atmosphere
USSR Voskhod 2	18 March 1965	Manned	1st spacewalk (for 10 minutes by Aleksei Leonov)
USSR Venera 3	16 November 1965	Unmanned	1st craft to land on another planet (Venus)
USSR Luna 9	31 January 1966	Unmanned	1st lunar soft landing; photographed surface
US Apollo 8	21 December 1968	Manned	1st manned orbit of the Moon
US Apollo 11	16 July 1969	Manned	1st landing on the Moon (Neil A. Armstrong and Edwin E. Aldrin)
USSR Luna 16	12 September 1970	Unmanned	Lunar soft landing; automatically took soil samples
USSR Soyuz 11	6 June 1971	Manned	1st inhabitation of Salyut space station (23 days)
US Pioneer 10	2 March 1972	Unmanned	1st satellite to relay data on Jupiter and leave Solar System
US Apollo 17	7 December 1972	Manned	Last and longest stay on the Moon (75 hours)
US Pioneer 11	6 April 1973	Unmanned	1st spacecraft enroute to Saturn
US Skylab 2	25 May 1973	Manned	1st occupation of Skylab space station
US Mariner 10	3 November 1973	Unmanned	Relayed 1st close-up photographs of Mercury
Soyuz Apollo	17 July 1975	Manned	Russian American link-up in space
US Viking 1	20 August 1975	Unmanned	1st successful landing on Mars
USSR Venera 9	8 June 1975	Unmanned	First photographs of Venus surface
US Voyager 2	20 August 1977	Unmanned	First spacecraft en route to Uranus and Neptune
US Voyager 1	5 September 1977	Unmanned	First detailed photographs of Jupiter, Saturn and their moons
USSR Soyuz 26	10 December 1977	Manned	First crew of long-term space station Salyut 6; first to be replenished by unmanned transporter; first to receive visitors
US Pioneer Venus	8 August 1978	Unmanned	First mapping of Venus topography by radar
USSR Soyuz 35	10 April 1980	Manned	Longest duration manned spaceflight (185 days) in Salyut 6
US Space Shuttle Columbia	12 April 1981	Manned	First re-usable space launcher

6 Medicine & Psychology



An introduction to illness and health

The men who inhabited the earth 20,000 years ago lived the life of the hunter, running after their prey and attacking it with stone axes. At the mercy of the weather, they took shelter in caves at night and protected themselves as best they could. For this life, men were as well, if not better, adapted than any other mammals, with bodies designed to endure long periods without food, to stand up to the heavy exertion of hunting and to respond quickly to the numerous emergencies with which their lives were fraught

The dangers of comfort

The physical characteristics of modern man are little different from those of his forebears but are deployed in very different circumstances. Food is plentiful, exercise is rarely taken and tensions in an urban society are continually high Early man's appreciation of sweet-tasting berries assured him of an adequate supply of water-soluble vitamins. The same craving for sweet things means that modern man's body is loaded with sugars and other carbohydrates, often in excess of his needs or his body's capacity to absorb them His sedentary, stressful urban life makes him overweight and tense In order to feel better he may smoke cigarettes, drink alcohol or take tranquillizers: palliatives that must, in the long term or in excess, do him harm The ingenuity of man's technology means that his body, which was designed to withstand the privation of a primitive existence, now has to function in a world that has been transformed by his intellect The body is ill-equipped to deal with the results of that transformation

The major health problems of developed societies stem from exactly this discrepancy between man's social and physical evolution In industrially advanced countries most of the diseases with easily identifiable causes, such as tuberculosis, smallpox or cholera, have been virtually eliminated and the diseases that increasingly matter are those caused by what may simply be described as an unhealthy existence High blood cholesterol levels resulting from an unbalanced diet, high blood pressure resulting from stress, obesity and smoking all add considerably to the chances of the occurrence of heart disease

Health is difficult to define It might be described as the subjective assessment of a person's ability to cope with his life: it is only when he feels ill that it is time to visit a doctor This definition, however, ignores the fact that some diseases do not produce any noticeable symptoms at their onset and so the tendency in industrial societies is to screen people for diseases before they actually feel unwell so as to prevent their ever getting seriously ill

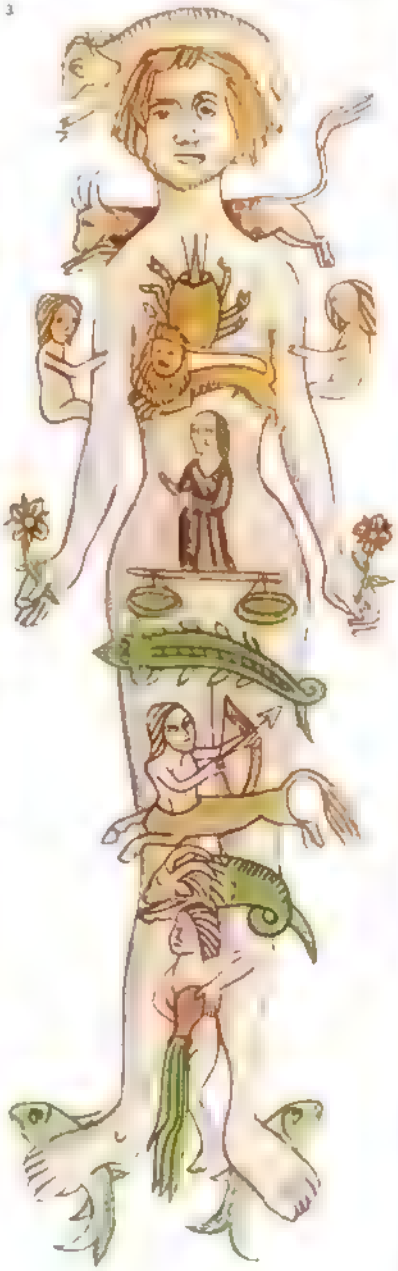
Disease prevention and detection

It is obvious that health checks help to reduce the incidence of epidemic diseases, but such monitoring also helps to reduce mortality from more general causes The first signs of the conditions that lead to many forms of heart disease or cancer – both scourges of developed societies – can be detected and steps taken to remedy the situation Studies in the United States have shown that the mortality rate among those who have check-ups is lower than normal and that it declines proportionately with the number of check-ups that the patient undergoes. A campaign conducted in the United States to increase public awareness of health hazards now

CONNECTIONS

1 Hippocrates, the father of medicine was born on the island of Kos c 460 BC He worked empirically, basing his judgments on observation rather than preconceived ideas The Hippocratic Collection, including the Hippocratic Oath that binds doctors to keep their patients' confidences, are medical works by many authors and are not necessarily of Hippocratic origin

2 Galen who was born in Pergamum in AD 130 was both an anatomist and a physiologist He proposed the theory that temperament was controlled by the balance of the four humours in the body (blood, phlegm and yellow and black bile)



3 Medicine and astrology were closely linked until the eighteenth century An astrological chart would be drawn up for the patient and used in diagnosis and prescription Each sign of the Zodiac is associated with a part of the body, so birth signs indicate which illnesses the subject is prey to Astrological relationships are also now suggested between the Zodiac and the glandular and nervous systems Aries rules the head and so Arians are said to be susceptible to headaches. Taurus rules the neck and throat, thus making Taureans vulnerable to colds The arm, shoulder and lung region is ruled by Gemini, while Leo rules heart, spine and back Cancer rules the stomach and chest making Cancerians prone to indigestion and chills Virgo rules the intestines and the nervous system, with the kidneys under the influence of Libra Scorpio rules the sex organs and Scorpians are the most highly sexed of all the signs Knees bones and teeth are ruled by Capricorn, so those under this sign are likely to be troubled by dental and orthopaedic problems Sagittarius rules liver, hips and thighs and Sagittarian women tend to have heavy thighs and hips Pisceans are prone to trouble with their feet and Aquarians will probably suffer from varicose veins and hardening of the arteries because Aquarius is the sign ruling circulation

4 Paracelsus (c 1493 1541), a Swiss physician and alchemist broke with the traditions of Galen and revolutionized medical methods Expelled from his post at Basel University in 1528, he continued with pharmaceutical experiments and encouraged research and development He started the treatment of venereal disease by mercury and believed that the physician should be alchemist, astrologer and theologian in order to tend the body, soul and spirit

seems to be contributing to the declining rate of fatal heart disease

In contrast to those who believe in check ups, some maintain that disease starts when the patient complains of symptoms and that over-anxiousness about health may actually lead to illness. A recent study, in which more than 25,000 English people were examined for diabetes, established that there were more undetected diabetics in the sample than there were known ones and that many of these people were symptom-free.

Another fact in favour of the argument that people are as healthy as they feel is that illness and mental strain are closely linked Coronary thrombosis is clearly related to a stressful life-style and is often precipitated by a worrying event The relationship between stress and illness has been so well defined that it is now possible to calibrate the chances of illness. Each event in a person's life is given a certain score highest on the list are deaths within a family, followed by deaths of close relatives changes in the home situation court appearances, holidays, conferences and so on A vulnerability score is obtained from adding up all these points. Recent studies of sailors have shown that the assessment of their vulnerability to illness accorded well with the incidence of illness during voyages

Fashions in Illness

Medicine, as all human institutions, is subject to changing fashion, both in the type and symptoms of disease and in treatment Before World War I it was fashionable for people suffering from mild psychological illness to display hysterical behaviour and no lady's handbag was complete without a bottle of smelling salts to cure a fit of the vapours Today's disease is depression; the symptoms are a slowing down and withdrawal from the situation and the remedy is the anti depressant tablet or tranquillizer. In the 1940s most children had their tonsils removed but tonsillectomy is now seldom performed unless medically necessary. Medical fashion varies not only with the times but from country to country. Circumcision, for example, is an operation widely performed in North America, whereas in Europe the trend has now moved against it

KEY

Egyptian medicine, which was largely based on magic, relied on diet, hygiene and emetics (as prescribed by temple priests) for its remedies Here Khemu fashions a man on a potter's wheel while his headed Thoth fixes his life-span

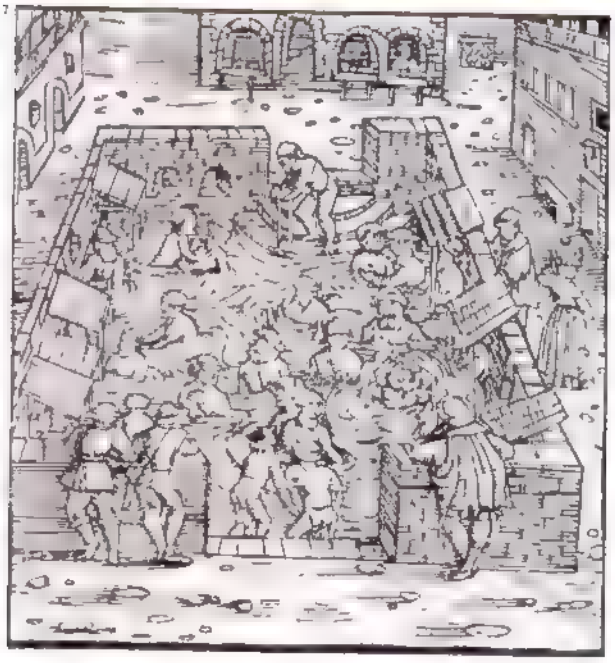
5

5 Leeches, collected from rivers by women who stood in the water and waited for the leeches to attach themselves, were used for blood-letting until recently Leech salivary glands contain an anticoagulant, hirudin, which has been used to prevent blood clotting, but has now been superseded by new drugs

6 Plethora, or excess of blood, was believed by Erasistratus (an early teacher in the Alexandrine School of Medicine) to cause many diseases He actually diminished blood by dietary methods, but his colleagues used blood letting (phlebotomy) widely and so began a practice lasting for many centuries

6

7



7 Mineral springs are to be found in almost every country and many cultures have exploited their curative properties. The Romans used hot springs as did the Crusaders, but spas were most popular in the 1700s and 1800s Places such as Baden Baden, Aix-les-Bains, Carlsbad and Bath became fashionable meeting places for the rich

8 Franz Mesmer (1734–1815), Austrian physician and mystic was the first man to use hypnosis and had remarkable successes with hysterical patients He was convinced that cures were due to his "animal magnetism" His theories prompted charlatans to produce devices, such as this tub, which were claimed to cure all illnesses

8

The causes of illness: 1

Few people today believe that "evil spirits" are the cause of epilepsy, that those who breathe marshland mists will develop malaria, or that smelly drains lead inevitably to an outbreak of typhoid fever. These ideas quite firmly held no more than a century ago have been almost completely dispelled by advances in modern medical science.

The scientific study of the causes of disease is known as aetiology. So successful has this science been that today doctors recognize thousands of disorders, and the discovery of new ones is still a fairly common event. Conveniently, however, nearly all the known causes can be grouped in a few general categories. For example, all the viruses, bacteria, protozoans and worms that cause disease are included in the category of infectious causes, and all the consequences of injury are grouped in a category called traumatic conditions [3].

Congenital disorders

Congenital causes of disorders are traditionally considered first for they act within the womb and cause disorders which are usually obvious at birth. Hare-lip, cleft-palate and club-foot are examples of common congenital disorders. Mongolism [1], deformities of the heart, some of which are responsible for "blue babies", and abnormalities of the nervous system associated with spastic paraplegia are others. As a class they are caused either by some fault in the chromosomal structure of the fertilized egg or by damage inflicted on the developing embryo in the womb – it is not always easy to decide which.

A developing embryo in the womb can be damaged by diseases contracted by the mother. German measles contracted during the early months of pregnancy, for example, can cause abnormalities of heart and ears (and sometimes of mentality and sight). Other congenital diseases are caused by drugs taken during pregnancy, such as the deformed babies that were born to mothers who had taken thalidomide. The smoking of more than ten cigarettes a day by a pregnant woman may retard the normal growth of the fetus in the womb.

The disorders that are transmitted from the parents to the child from generation to generation are known as hereditary disorders. Perhaps the most well-known is haemophilia, a disorder in which the clotting mechanism of the blood is deranged so that even trivial injury is followed by prolonged bleeding. Another hereditary condition is phenylketonuria – in which there is an absence of enzymes that metabolize certain toxic phenyls. Untreated, it leads to severe mental retardation, but if detected soon after birth it can be dealt with quite successfully by a diet that is low in the amino acids that give rise to phenyls in the bloodstream.

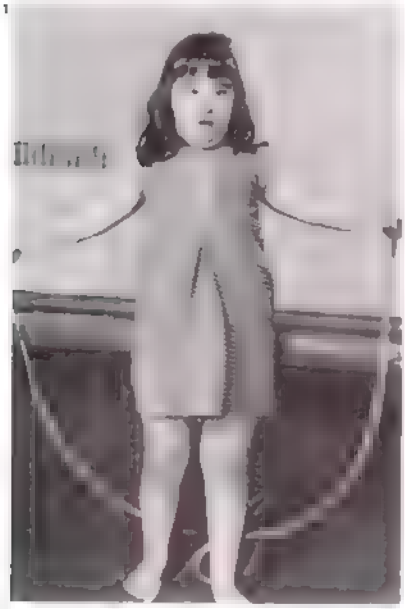
Dietary deficiency

Starvation and malnutrition are obvious causes of disease. Even so, a seemingly ample diet may cause disease if it lacks certain vitamins, the nutrients essential for well-being. Vitamins occur only in certain foods and a diet that consistently lacks a sufficient quantity of a particular vitamin is certain, in time, to give rise to the corresponding vitamin deficiency disease [2]. The scurvy that afflicted sailors on long sea voyages and the beri-beri

CONNECTIONS

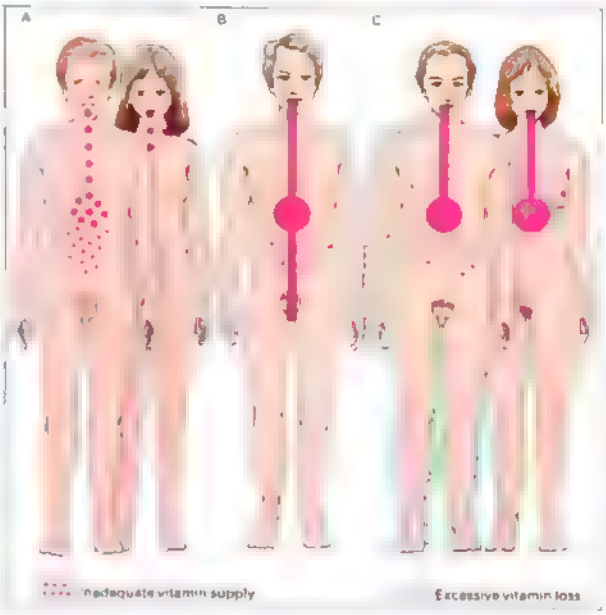
See also

1



1 Mongolism, a form of mental deficiency associated with certain physical characteristics, is caused by a chromosome imbalance present at birth. Normal cells have 46 chromosomes whereas mongoloids either have 47 or one of the 46 may be oversized. The excessive chromosome material interferes with the normal control of brain and body growth, producing the flattened face and folded upper eyelids that give the patient an Oriental look. Other features are short arms and legs, thick stubby hands and a large abdomen. The condition, also known as Down's syndrome, is more likely to occur in babies born to older women.

2



2 Vitamins are essential to a healthy diet and lack of them causes a variety of deficiency diseases such as rickets (vitamin D deficiency), scurvy (vitamin C) and pellagra (niacin). Shortage of vitamins can be due to a primary food shortage or to diminished food intake [A] resulting from poverty or taboos against eating certain foods (beef for Hindus and pork for Jews and Muslims). Even when a diet is rich in vitamins, parasitic infections or impaired transport in blood or lymph vessels can reduce absorption [B]. Great physical activity, growth, or excessive loss through sweating or breast feeding [C] create the need for an increased vitamin intake.

3

3 Accidents are listed under the general heading of traumatic conditions. They range from simple cuts, burns and scalds to the direct effects of a road accident. In the same category doctors put injuries caused by lightning, electrocution, sun burn from ultra-violet light, frostbite and the air pressure problems of aircraft pilots or divers. Exposure to radiation from X-ray machines, radioisotopes or atomic explosions destroys various body cells. Some cells are more easily injured than others. Bone marrow, where red blood cells are produced, is most sensitive, then the lining of the stomach and intestines, the skin and sex glands and lastly cells of the brain and muscles.

from which thousands of prisoners in Asia died during World War II were caused in this way. The childhood disease of rickets is also caused by vitamin deficiency. In the industrial countries of the world, vitamin deficiency diseases are rare today but occasionally occur in people who, through preference or ignorance, choose to live on very restricted diets. Strict vegetarians, for example, are liable to suffer from lack of vitamins A, D and E which are contained in animal fats.

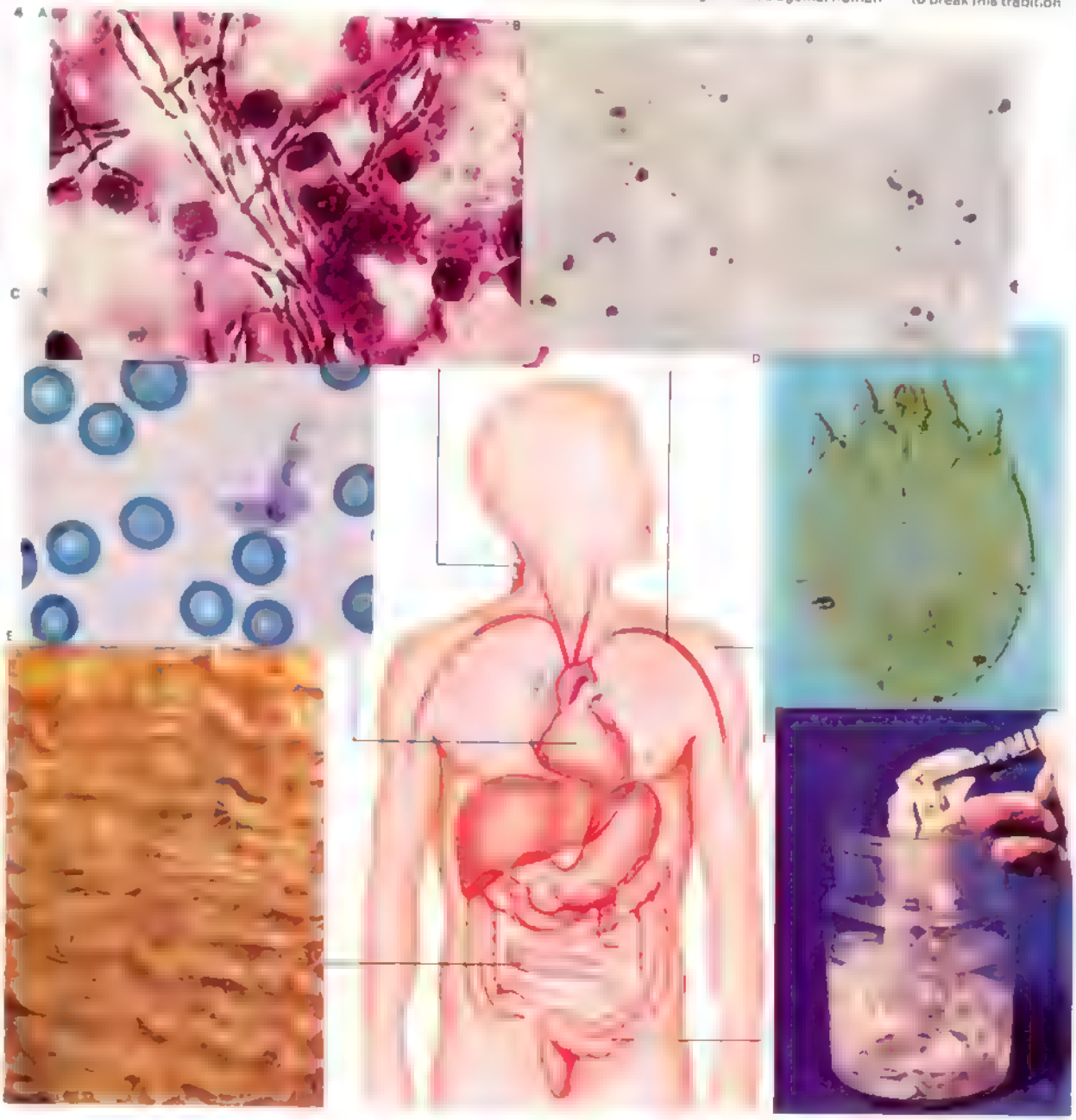
It is not only a lack of vitamins, however, that may cause deficiency diseases. There are several other constituents of the diet that are essential for health, and important among these is iron. A little of this mineral is needed to make the haemoglobin in the blood's red cells. A deficiency of iron causes anaemia. Another element needed by the body in small amounts is iodine, a shortage of which leads to an enlargement of the thyroid gland, a condition known as goitre.

Infections and their causes

An enormous number of living things gain entry to the body's tissues, grow there, and by so doing cause disease [4]. These infectious agents range in size from viruses so small that they can be seen only with the aid of an electron microscope, to bacteria and protozoa that can be seen with an ordinary light microscope and the various small creatures just visible to the naked eye, to the tapeworms which grow to several feet.

The way in which infectious agents cause disease is extremely varied. Viruses are intracellular parasites and the polio virus, for example, causes paralysis by growing in and destroying a particular type of nerve cell in the spinal cord. Bacteria, on the other hand, are largely extracellular and cause their ill effects either by secreting the powerful poisons known as bacterial toxins or by invading tissue. The malaria parasite destroys the blood's red cells and the amoebae of amoebic dysentry irritate and poison the bowel. The effects of infection or infestation with larger parasites are less well defined but there is little doubt that the debility associated with this type of infection is largely due to interference with nutrients and loss of blood from the intestinal wall.

Anatomy classes based on thorough human dissection, like this one at the Barber Surgeons' Hall London, 1581, were the foundation of a proper understanding of how the healthy body worked, and consequently how diseases were caused. Previously such scientific knowledge was hampered by a strong taboo against human dissection and an over reliance on such classical writers as Galen (c. AD 130–c. 200). The Belgian Andreas Vesalius (1514–64) was the first to break this tradition.



4 Infectious diseases are caused by invading organisms. Microscopic ones (germs) include viruses and bacteria. Larger are protozoan and fungal infections such as athlete's foot (tinea or ringworm). Larger still are flatworms and roundworms. Viruses cause many common diseases such as mumps, measles, chicken pox, influenza and the common cold. Some bacteria can live "dormant" in the body until fatigue or disease lowers resistance. Streptococci, for example, are often present but only sometimes cause a sore throat. Bacterial diseases include leprosy, tetanus, whooping cough and diphtheria. *Bacillus anthracis* [A] causes anthrax, an infection of animals which can be caught by those handling them. Symptoms are boils, fever and death in two or three days if untreated. Protozoa, minute one-celled parasites, cause many diseases. Among them are *Plasmodium*, responsible for malaria [B], which appears as minute "signet rings" in the red blood cells, and the trypanosome [C] that causes sleeping sickness. The mite *Sarcoptes scabiei* [D] burrows into the skin causing intense irritation. Undercooked infected pork and beef are the cause of worm infection. *Trichinella spiralis* [E] lives harmlessly in the intestine but the larvae migrate to muscle where they form small painful cysts. The tapeworms *Taenia solium* and *Taenia saginata* [F] are man's largest parasites.

The causes of illness: 2

Many of the agents that infect human bodies produce poisonous substances but there are many poisons that originate outside the body and that are equally damaging [1]. Many of the chemicals used in industry are poisonous, most drugs are poisonous if taken in excess, many plants synthesize poisons in their leaves and fruits, and numerous animals, including some snakes and spiders, use venoms either for attack or defence.

Despite their diverse origins poisons have a marked similarity of action for nearly all of them interfere with one or more of the chemical reactions that take place in living tissues. As a consequence of this interference the poisoned tissue ceases to function properly and, in severe poisoning, may die.

Some poisons (the kind favoured by murderers and those attempting suicide) lead rapidly to death rather than to illness. Potassium cyanide is probably the best-known rapid poison and it works almost instantaneously by interfering with the intracellular oxidations responsible for the production of the energy used in all living tissues. Any cell it meets stops functioning immediately.

Carbon monoxide rapidly unites so strongly with haemoglobin that it prevents the blood from taking up any oxygen and swiftly results in the suffocation of the body's cells. Less immediately lethal poisons are those likely to be responsible for lasting illness. These slower poisons can affect many different chemical reactions and so produce different symptoms. Lead poisoning is cumulative, slowly disrupting the formation of red cells.

Tumours and cancers

Neoplastic (new growth) diseases are those that are associated with the growth of tumours. Generally speaking it is the extra growth that brings on the symptoms and so as the underlying causes for the tumour are often not known [4], they are classified according to the size and site of origin.

Tumours vary from harmless warts commonly seen on hands and fingers to fatal cancers. A large proportion are slow growing with little tendency to spread to other parts of the body. This kind of tumour causes disease merely by its presence, and its position is more important than its size [5]. A large tumour under the skin of the back may be harmless while a quite small one, in contrast, can cause disabling disease if it presses on an important or delicate structure such as a nerve and so deranges that structure's function. An acoustic nerve tumour, for example, is harmless in itself but causes deafness.

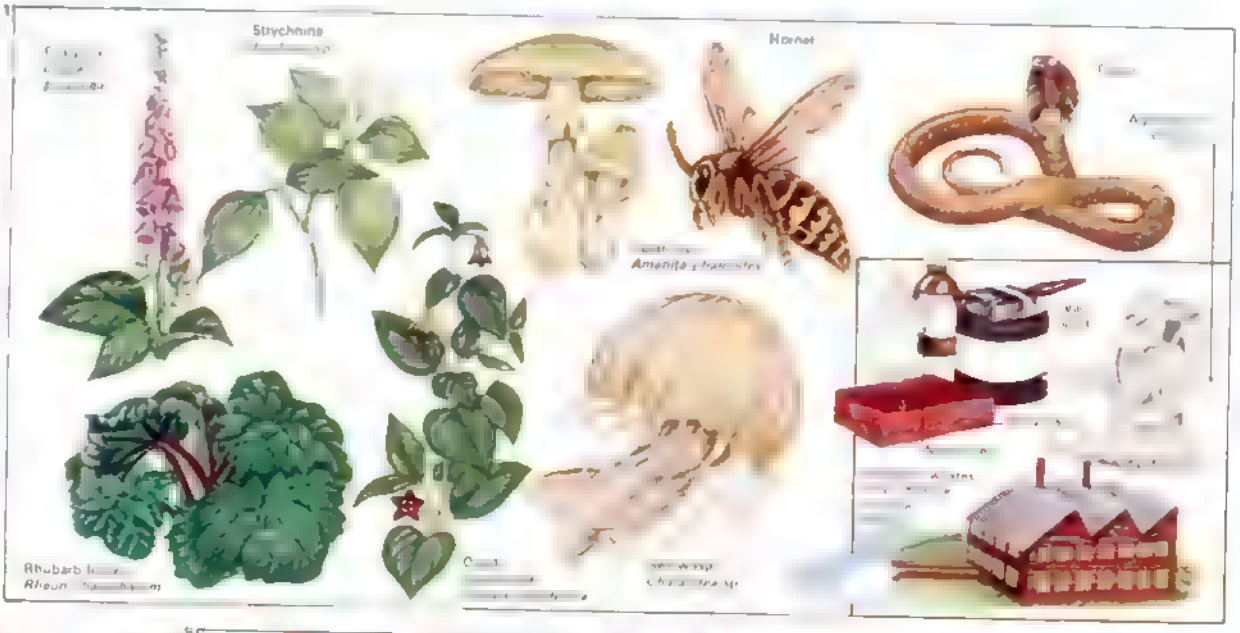
In contrast to the slow-growing, non-spreading tumours are those that grow rapidly and are capable of spreading to almost any part of the body. These are the cancers. A tumour of this kind causes disease not merely by its destructive effects on the tissue in which it originates and grows, but also by similar effects on any tissue to which it spreads. It is because cancers grow and spread so rapidly that they produce such serious symptoms and are so dangerous.

Degenerative and immunological diseases

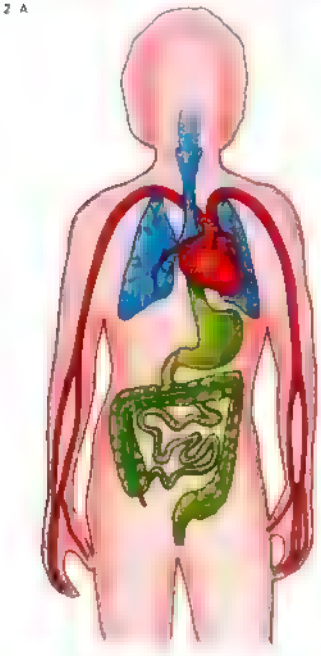
Human tissues, like machines, suffer from wear and tear but, unlike machines, they are able in many cases to repair themselves. With increasing age, however, the reparative ability declines and as damage outstrips

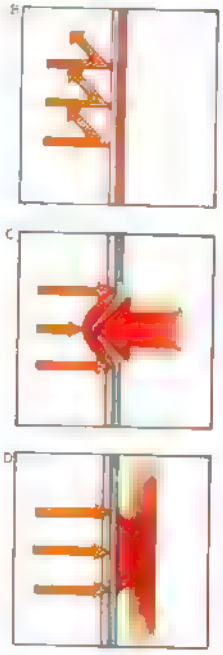
CONNECTIONS

See also

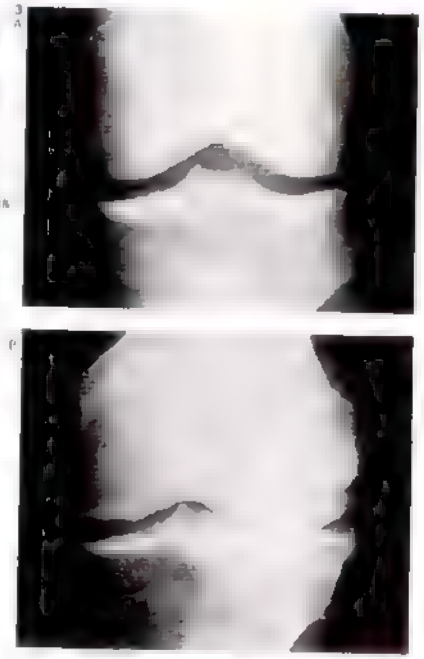


1 Poisons may be inhaled, injected, absorbed through the skin or swallowed; a few common ones are shown here. Some, such as hornet stings, are mild although even these may kill, others so toxic that they can kill within seconds. Most deaths, however, are caused by weaker poisons in household goods. Children are particularly at risk since they are more sensitive to poisons than adults. The most difficult to avoid are those that occur undetected, for example, the bacterial toxins in contaminated stews and pâtés. In special circumstances harmless substances are poisonous: cheese can be lethal to anyone taking certain anti-depressant drugs.





2 Allergies [A] can be caused by almost any substance. Pollen and dust affect the respiratory system [blue]. Some foods upset the digestive system [green]. Drugs, either injected into the blood stream [red] or eaten, cause general reactions. Many substances on contact with the skin cause local skin reactions [pink]. Potent allergens [arrows] have no effect on non-sensitive people [B]. Cells of sensitized people are coated with antibodies specific to the allergen [green layer]. Contact with the allergen causes these cells to release histamine which causes inflammation [C]. Antihistamine pills, sprays or injections suppress allergic reactions by blocking histamine action [orange layer, D].



3 Osteoarthritis, a painful degenerative condition of the joints, is a common affliction of the elderly. Comparison of X-ray pictures of a normal knee joint [A] and an arthritic one [B] shows the characteristic loss of 'joint space'. The large joints at the knee, hip and shoulder are most commonly affected but small joints like those of the hand are often involved. The treatment consists mainly of less activity, heat therapy and pain killers. In severe cases cures include injections of cortisone drugs, replacing the joint with a new plastic one, or even ankylosis, which involves removing any remaining cartilage and immobilizing the joint until the two bones fuse.

repair the symptoms of degenerative disease appear. Most common of all degenerative conditions are greying of the hair, baldness and the loss of elasticity of the lens of the eye which most middle-aged people suffer from. Far more serious, however, is the painful osteoarthritis that so commonly develops as the surfaces of the joints of elderly people begin to wear [3] and several degenerative conditions of the nervous system, such as Parkinson's disease.

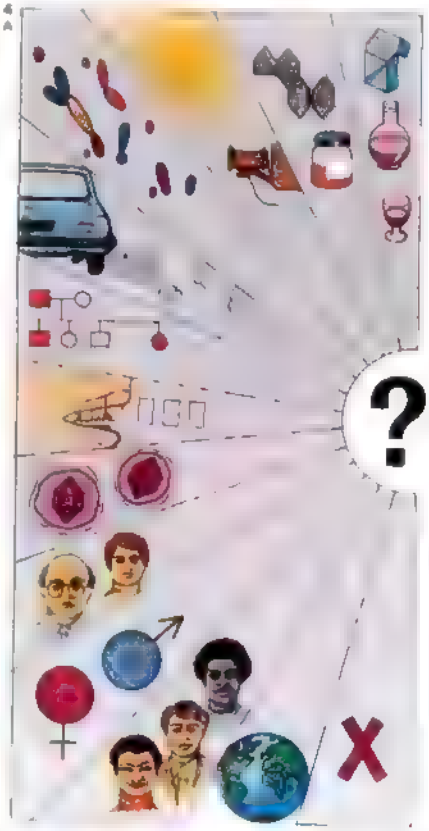
Hardening of the arteries or arteriosclerosis, when the arteries lose their smoothness and elasticity, leading to angina pectoris, coronary thrombosis and stroke, is the most important of the degenerative diseases in industrialized countries, where it is responsible for more deaths than any other condition.

The immune defences of the body provide protection against infections and probably against the onset of many cancers. Sometimes, however, the defences become directed against quite innocuous "enemies" or even against the body's own tissues. Abnormal responses such as these are responsible for hay fever, urticaria (nettle rash) and asthma [2] and also for the far more serious conditions of rheumatoid arthritis and haemolytic anaemia. The immune reactions that cause these diseases are similar to those responsible for the rejection of kidney grafts and heart transplants. And occasionally mothers respond to the red blood cells of their unborn infants by producing antibodies that cross the placenta, damage the infant's blood and result in the child becoming what is commonly known as a Rhesus baby.

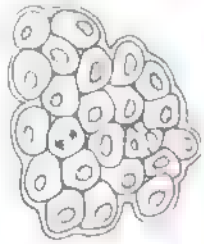
Iatrogenic and idiopathic diseases

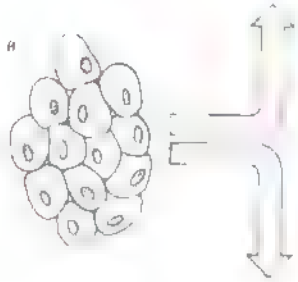
The final categories of diseases known to doctors are the iatrogenic diseases and idiopathic diseases. Iatrogenic diseases are those that are a consequence of medical care and range in severity from the acceptable side-effects of a drug necessarily given in the treatment of a serious illness to the disasters that sometimes follow accidents in operating theatres. Idiopathic diseases are those for which no cause has yet been found, such as the skin disease psoriasis. For this reason they pose great difficulties in treatment.

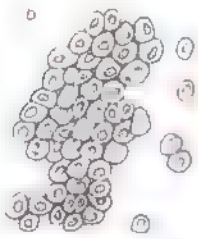
Cancer is a major killer particularly in countries with a high life expectancy. Research is currently centred on viruses, for more than 100 viruses have been identified as carcinogens, or external cancer causing agents in animals. Positive proof, however, of the ability to cause cancer in man has yet to be fully demonstrated although some researchers have associated herpes virus with cancer of the cervix.



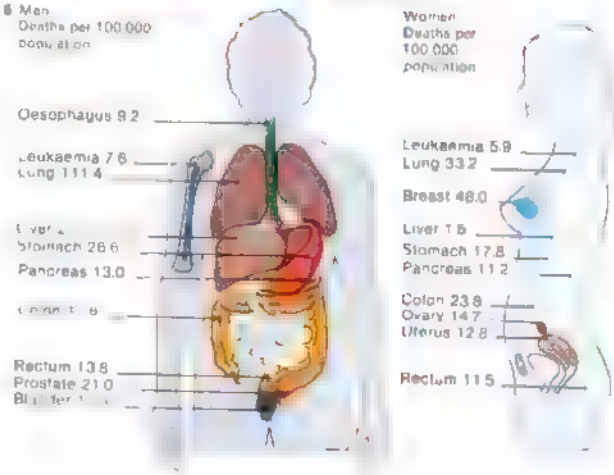
4 The mechanisms by which normal cells mutate into cancer cells are not known but several carcinogens, external cancer causing agents, and other factors that may have an effect have been identified. [A] These include various chemicals such as those released by cigarettes, hormones, radiation, chronic long-standing infection, air pollution, genetic factors (certain types of cancers run in families), chronic irritation or abrasion as caused in the mouth by a jagged tooth, viruses, age, sex (lung cancer is more common in men), race and geographical location. Normal cells have a specialized structure, appearance and function [B]. In benign tumours [C] this is maintained to an extent but in malignant tumours [D] cells are often grossly aberrant. They also lose their natural ability to stick together and so break away and spread the cancer.







5 Lipoma is one of the most common of the slow-growing non-cancerous tumours. This tumour consists of large numbers of fat cells that grow in such a way that they form a mass, often just beneath the skin. Such tumours can be unsightly but they are seldom dangerous. Surgical removal is usually feasible, especially when the tumour is small as here and this will probably prevent any further growth.



6 Cancer causes death by physical obstruction of an essential pipeline, by pressure on normal tissues (such as within the skull) and by crowding of normal tissue. This results in tissue death and consequent loss of the function performed by that tissue. Mortality is shown for the ten most frequent sites of cancer in England and Wales during 1980. As such, the figures highlight the cancer situation only in relatively affluent societies.

7 Cancerous tumours of the skin are called epitheliomas because skin is an epithelial tissue. Such tumours occur with greatest frequency in particular sites such as the face, lips, ears and penis. The rapidly growing cells of the epithelioma spread readily to adjacent lymph nodes and for this reason it is important that these tumours are diagnosed and treated early. They often begin as small sores that refuse to heal.

Diseases of breathing

The agents that infect the respiratory passages and the lungs [Key] are basically similar to those that cause disease in other systems. Ports of entry into the body for food, water or air – the mouth and nose – are also potential sites of weakness in the body's defences. The lining of the respiratory passages is especially vulnerable to various infections caused by invading micro-organisms.

The symptoms of infection

The tonsils and adenoids are prominent parts of the lymph system in the throat and help repel infection. Invasion by viruses or bacteria may result in sufficient inflammation to give tonsilitis [1] in which the throat and neck become sore and swallowing is painful. Many similar viruses of the rhinovirus group can affect the throat, nose and eyes to give "the common cold". Infections of the respiratory passages are named according to the part most affected – larynx (laryngitis), throat (pharyngitis), windpipe (trachetis) and lung tubes (bronchitis)

When lung tissue itself becomes infected pneumonia develops [3]. Various bacteria can invade the lung air sacs causing inflammation, increasing blood flow to damaged tissue and preventing oxygen from entering the blood in sufficient amounts. Toxins, poisonous substances produced by the microbes, further affect the patient. Before the development of antibiotics recovery from pneumonia depended on the ability of the patient's own defences to fight the illness and mortality was high. Today the administration of appropriate antibiotics means quicker recovery, providing that pneumonia is not secondary to another serious debilitating disease

Other bacterial infections of the lung tissues may not initially be as dramatic. Tuberculosis [6], once known as "consumption", may develop slowly before becoming obvious. Many children appear to come into contact with the disease and rapidly develop immunity. With careful treatment a cure can be expected with the drugs now available. Tuberculosis is more likely to occur in people not in the best of health – the elderly, those with chronic chest disease such as bronchitis and emphysema and in people with poor general nutrition such as chronic alcoholics

The first symptom of a lung disorder is usually interference with breathing. This may take the form of pain, breathlessness, phlegm or catarrh The breathlessness may appear suddenly or slowly, sometimes during rest or sometimes only on exertion

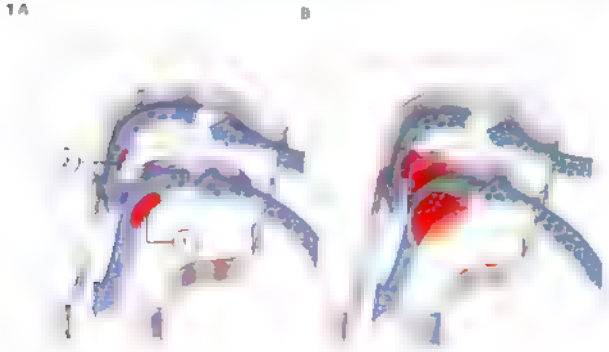
Irritants of the air passages

Two lung diseases, bronchitis and emphysema [2], impede the passage of air to and from the lungs. In bronchitis the blockage is often due to the accumulation of excessive mucus; in emphysema to the compression and collapse of air tubes caused by unnaturally distended lung air sacs or alveoli in which air has become trapped. Chronic bronchitis is caused by continual irritation of the bronchial mucosa often due to tobacco smoke, dust, smoke and fumes.

It is not surprising that, as the years pass, diseases of the lungs develop, for the respiratory system suffers a considerable degree of abuse. This may take the form of large quantities of fine dust particles that are present in all air, but are often worse in

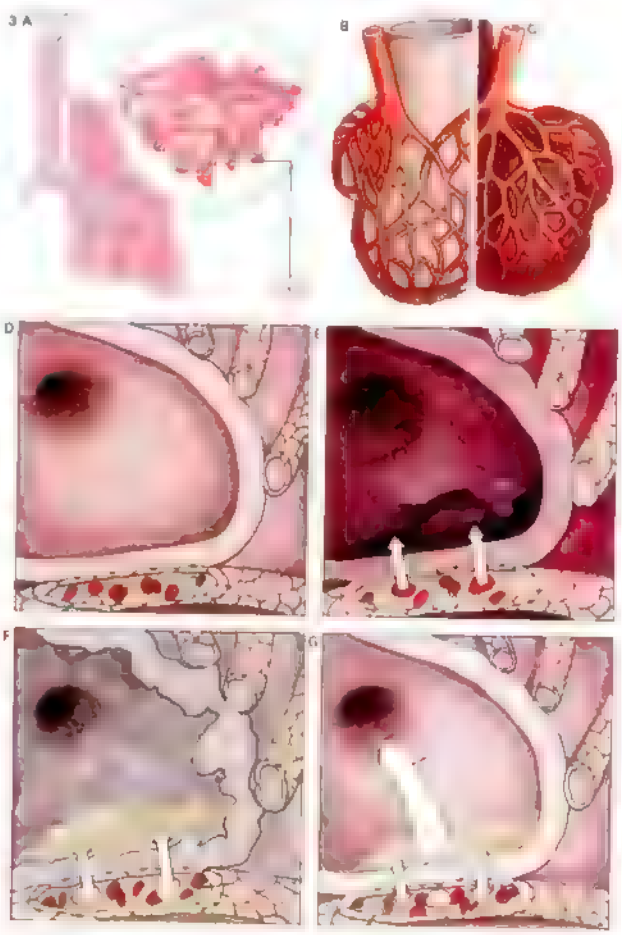
CONNECTIONS

See also



1 Air breathed in through the mouth and nose [A] carries with it viruses and bacteria The tonsils [1] at the back of the mouth and the adenoids [2] in a child are part of the body's defences against potentially infective microbes These organisms may lead to inflammation of the tonsils with painful swelling [B], an increased temperature and difficulty in swallowing The infection may require antibiotic treatment

2 Viruses in the air pass down the trachea (windpipe) into the lung tubes (bronchi) where they may cause asthma acute tracheitis and bronchitis with symptoms of cough and fever In asthma [A] the small air tubes become excessively narrow [1] causing difficulty in breathing The alveoli overfill [2] but the condition is generally intermittent Chronic inflammation of the tubes – called chronic bronchitis [B] occurs mainly in older men who work in industrial areas and who smoke A winter cough and wheezing occur, worsened by viral infections The inflamed lining of the bronchial tubes [3] produces excess mucus [4] which is coughed up as phlegm Inflammation and narrowing of the bronchi [5] lead to difficulty in moving air into and out of the alveoli [6] The distended alveoli [7] are damaged and some may collapse [8] The condition called emphysema [C] stops oxygen getting into the blood normally



3 Infection may not affect only the upper respiratory system and the lung tubes [A], but also the lung tissue and alveoli [1], resulting in pneumonia This is normally an acute infection by invading bacteria but may also be caused by viruses. Normal lung alveoli [B] become engorged with blood, inflamed and thickened [C] Sections through the alveoli seen under the microscope show the stages of the disease. Compared to the normal lung [D], tissue in a pneumonic lung [E] becomes inflamed and the alveoli are invaded by plasma and red and white corpuscles. These form a viscous mass that coagulates, causing acute congestion and airlessness. As the disease progresses [F] some tissue may die and excess of white corpuscles gives the lungs a grey appearance During the patient's recovery [G] the dead bacteria, fluid and inflammatory cells are either coughed up or reabsorbed into the blood stream. Treating the patient with antibiotics speeds recovery, while physiotherapy, by helping to ease breathing and by encouraging the coughing up of infected phlegm, minimizes permanent lung damage

industrial areas [4]. The lining of the nose, which is equipped with mucus-secreting cells and long hairs, efficiently traps and filters most dust. Not everyone breathes through the nose, but the mouth and upper air passages are also quite efficient traps

Lung disease can often be traced to a particular irritant; certain industrial processes produce silica particles that are a particular menace because they reach the alveoli in sufficient amounts to set up, over a period of time, chronic inflammation and scarring. This greatly reduces lung function and increases the risk of lung cancer. Other hazards are forms of asbestos which, when breathed in, increase the risk of cancer of the large air tubes (bronchi). These are also the major cause of a cancer of the pleura – the lining of the lungs and the chest wall Restriction of the use of such substances, and removal and prevention of coal-dust and silica-containing dusts, are important medical preventive measures

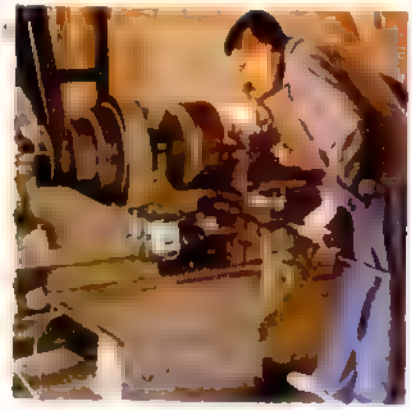
By far the greatest threat to the health of lung tissues is smoking [7]. Smoking increases the irritability of industrial dusts, is a major factor in lung cancer, greatly increases the risk of debilitating chronic bronchitis and emphysema and also has far reaching effects outside the lungs – for example, it increases the risk of heart disease

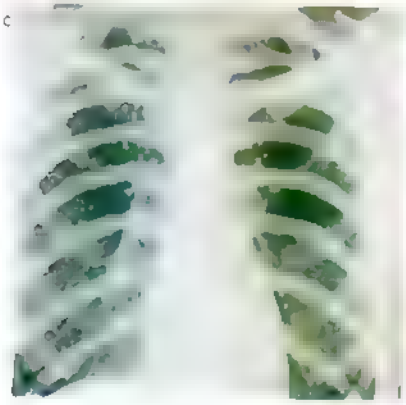
The incidence of different lung diseases varies from country to country and sometimes the reasons for this variation are obscure. Chronic bronchitis and emphysema, for example, have internationally been called the "English diseases", for they have a particularly high incidence in the United Kingdom More males are affected than females and these diseases are more common in towns

Allergies: asthma and hay fever

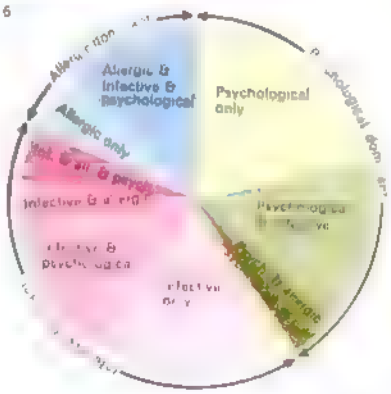
Other inhalants may produce allergic responses. In the summer months grass pollens, for example, produce hay fever whose symptoms are streaming noses and eyes, asthma [5] and general ill health in some people. Allergic responses can occur under conditions of prolonged exposure to a highly concentrated foreign material Thus farmers may suffer asthmatic-type reactions when filling or emptying silos

An oxygen supply is required for all the body's processes and is taken into the body through the lungs from which the waste gas carbon dioxide is breathed out The nose [1] and mouth [2 can both be used for breathing in and out The nose filters the air which passes to the windpipe [3], the larger and then the smaller bronchi before reaching the many million air sacs (alveoli) of the lungs [4] Many of the disease-causing agents enter the body through the same pathway Various traps *en route* help to deal with these injurious agents and second line defences are available to tackle any that get as far as causing d

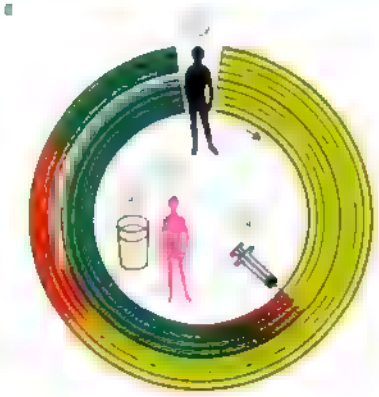




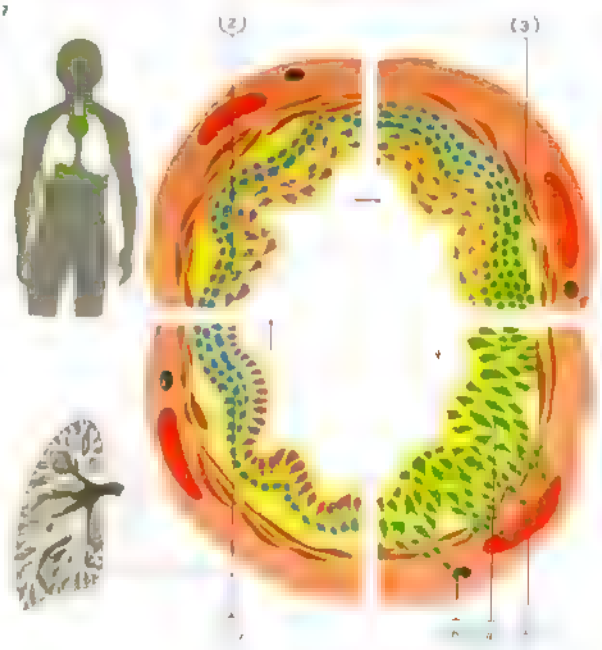
4 Industrial dusts can cause damage to the lungs Industries whose workers may be particularly at risk are coal mining [A] – and those in which fine silica particles occur in the air – granite and sandstone industries, the pottery industry and metal grinding processes [B] In the early stages there are no signs or symptoms but gradually breathlessness, a cough and phlegm develop The dust causes inflammation and scarring of the lungs and X-ray appearance changes from the normal [C] to one with patchy mottling and scarring in the lungs [D] Smoking speeds lung damage Industrial dust removal is an important preventive



5 An attack of asthma can be precipitated by a number of factors These may be psychological, infective or allergic, or occur in combination, varying in importance with individuals While one factor, such as psychological stress, may be most important, an attack becomes more likely if an infection or an allergy is also present The size of the segments indicates the relative importance of each cause or combination of causes



6 Tuberculosis normally affects the lung tissues but may spread throughout the body A healthy person [1] can become infected by inhaling bacteria spread by a tuberculous patient [2] or by drinking cow's milk [3] containing bovine tubercle bacilli A long, often chronic infection follows [red] with a 25 per cent mortality risk Those who recover become immune [green] Immunization [4] by modified bacilli [orange] will also give a person protection against tuberculosis



7 Lung cancer most commonly affects smokers. Cancers develop from cells that alter their structure and multiply more rapidly Changes may occur in cells of the inner or outer lining of the bronchial tubes [1] Altered malignant cells increase in numbers [2] and extend locally [3] or invade adjacent tissue [4] In addition the cancerous cells may spread outside the lung either through blood vessels [5] or through the lymphatic system [6]

Diseases of the circulation

An obvious feature of the body's circulation system is the close interdependence of the heart, the blood vessels (arteries, veins and capillaries) and the blood. Abnormality or disease of any part [Key] often leads to malfunction of the whole system.

Congenital malformation that disrupts the flow of blood through the heart and/or its major blood vessels may lead to a baby appearing blue soon after birth because the blood is carrying too little oxygen. An operation may be needed soon after the disorder is diagnosed or perhaps later in childhood to correct the abnormality.

Types of heart disease

The heart and its valves are susceptible to damage by disease – a commonly known one being rheumatic fever which frequently leads to rheumatic heart disease. The damage it causes to the valves in childhood can lead to severe malfunctioning of the heart in later life but in many cases a valve-replacement operation can provide a remedy.

Occasionally the heart's pacemaker beats irregularly, or a defect in the conduction of the electrical impulse that causes the beat may lead to a chamber beating out of sequence. These so-called dysrhythmias may be helped by drugs (such as digitalis) but in some cases an artificial pacemaker is needed [2].

Any decrease in the supply of blood to the heart itself through the coronary arteries can lead to attacks of pain, called angina pectoris, a kind of cramp of the heart muscle that is usually worse after exercise or effort. Severe restriction in the blood supply can cause failure of the heart muscle. This is accompanied by sudden excruciating pain in the chest and is commonly called a heart attack.

The usual cause of a heart attack is a thrombosis or clot in a coronary artery [1]. Definite causes of coronary thrombosis are still largely speculative. A person who is overweight, takes little exercise, smokes, has high blood pressure, a high level of cholesterol in the blood and a family history of heart disease incurs a high risk of attack.

Defects in arteries and veins

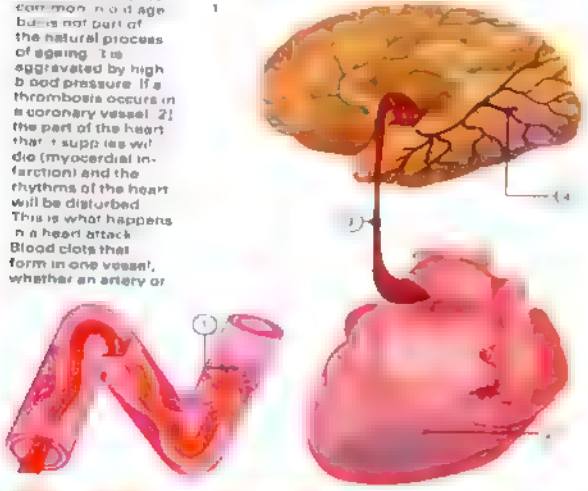
Hardening of the arteries (arteriosclerosis) is a degeneration of the wall of the artery and is a normal part of growing old. It affects all the arteries of the body and can be accelerated and exaggerated by high blood pressure. Atherosclerosis (deposition of fats and cholesterol within an artery) is much more serious. It causes narrow, roughened vessels and may lead to thrombosis. If it occurs in the arteries of the legs, exercise may become painful and a graft may be needed to replace the damaged artery. Atherosclerosis of the aorta may reduce the elasticity of the vessel, producing a bulging of the wall called an aneurysm, especially in the lower back. This may rupture spontaneously and cause severe haemorrhage needing emergency treatment.

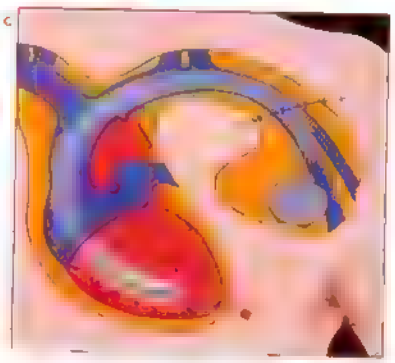
In many cases, the cause of high blood pressure or hypertension is unknown but hardening of the small arteries of the kidney is certainly contributory. As high blood pressure makes this condition worse, treatment is needed to break the vicious circle. Several drugs can help. Some act directly to remove adrenaline (which causes the arteries to contract). Others tranquillize a patient and so reduce the amount of adrenaline he produces. Digitalis helps the heart to pump more

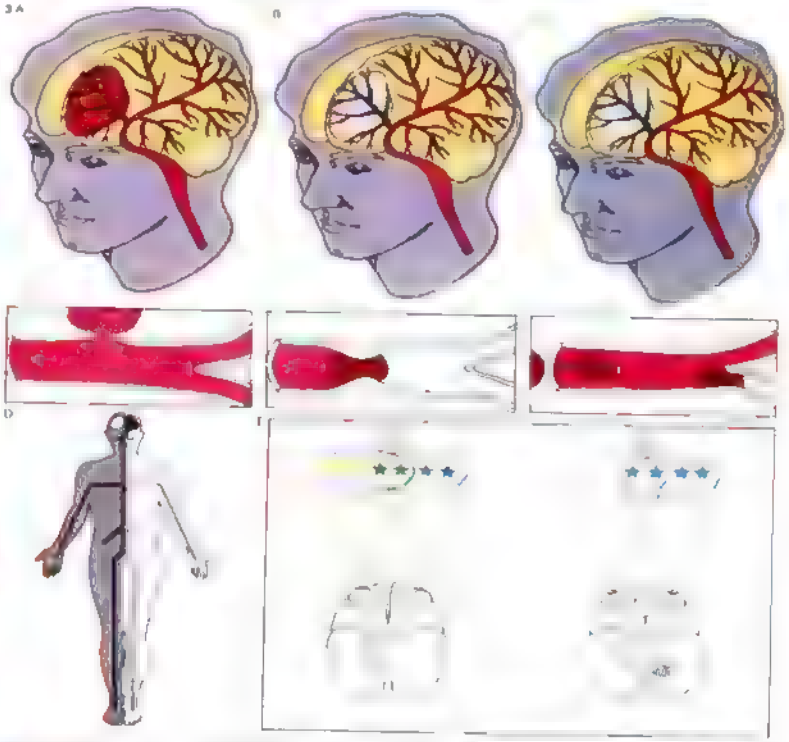
CONNECTIONS

See also

1 Thrombosis is the blocking of a blood vessel by a blood clot. [1] This commonly happens in an artery that is already narrowed and roughened by fatty deposits, a condition called atherosclerosis, which is quite common in old age but is not part of the natural process of ageing. It is aggravated by high blood pressure. If a thrombosis occurs in a coronary vessel [2] the part of the heart that it supplies will die (myocardial infarction) and the rhythms of the heart will be disturbed. This is what happens in a heart attack. Blood clots that form in one vessel, whether an artery or vein, may break off and then travel [3] to lodge in another vessel. A clot of this kind is called an embolus; if this occurs in cerebral vessels [4], it can deprive the brain cells of oxygen and thus cause a stroke. In a few cases, surgery to the damaged vessel may be helpful but physiotherapy and rehabilitation are the mainstays of treatment in most cases of this sort.







2 The heart has a natural pacemaker (A), the sinoatrial node [1], which controls heartbeat and rhythm. It transmits regular impulses through special conducting tissues, stimulating heart muscle contraction. Defective conducting tissue may result in atria and ventricles contracting in a dissociated way which may lead to complete heart block (B). To avoid this possibility, a battery operated artificial heart pacemaker can be used to stimulate ventricular contraction at a normal rate of 70 to 80 beats per minute. The pacemaker can be placed under the skin of the chest with the electrode leads passing through a great vein to the ventricle (C). Batteries for pacemakers are changed every two years.

3 A "stroke" is a disruption of the blood supply to the brain and is due either to a haemorrhage or a cerebral blood vessel. At a thrombus (B) or an embolus (C). The damage (stippled area) may be permanent but the outcome of a stroke depends on the extent and area of the brain affected by it. The symptoms can range from a temporary loss of speech or other brain function and paralysis of the limbs to sudden death. A stroke on one side of the brain affects the limbs of the opposite side of the body because of the crossing over of nerve tracts in the brain stem. D] There is a similar effect on the visual cortex. E] Normal vision [1] is impaired on the left side. 2] by a stroke on the right side.

vigorously while drugs called diuretics reduce the fluid in the circulation

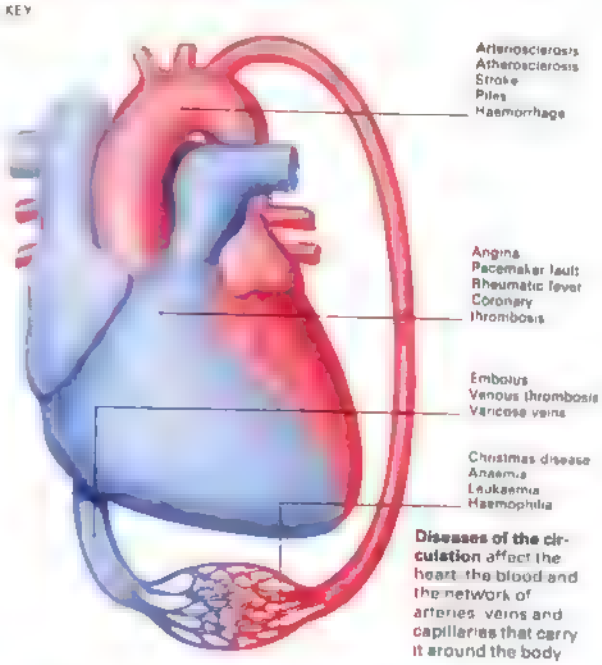
Piles and varicose veins [4] are common conditions, both made worse by long periods of standing. Clots can also develop, especially in the legs, when the veins are inflamed (phlebitis). The leg will be swollen and painful but in plain thrombosis there may be no signs and diagnosis is difficult. Venous thrombosis is liable to occur in people who have prolonged periods in bed after an operation or a stroke. The danger of a clot (which can be 15-18cm [6-7in] long and 0.5cm [0.2in] wide) in the legs is that it may break off to become an embolus which travels through the heart and lodges in the lungs. Thrombosis is usually treated with anticoagulant drugs

Blood defects

Blood is subject to a number of diseases. In anaemia there is usually either a reduction in the number of red cells or in the cell's haemoglobin content [5]. The reduction in the oxygen-carrying capacity of the blood causes tiredness and breathlessness. More rarely, disorders of the bone marrow which makes the blood cells lead to anaemia, sometimes the red cells are destroyed more quickly than their usual four-month life-span and are not replaced fast enough

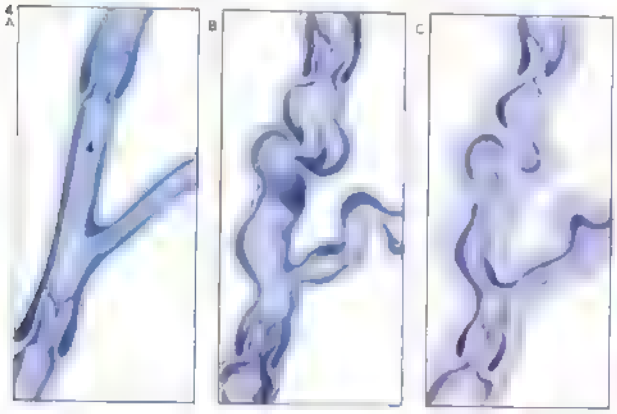
Leukaemia [8] is the name given to a group of diseases characterized by proliferation of abnormal white cells. The patient is likely to suffer from anaemia, infection and bleeding. Drugs, including steroids, that interfere with the reproduction of white cells are used in the treatment of leukaemia

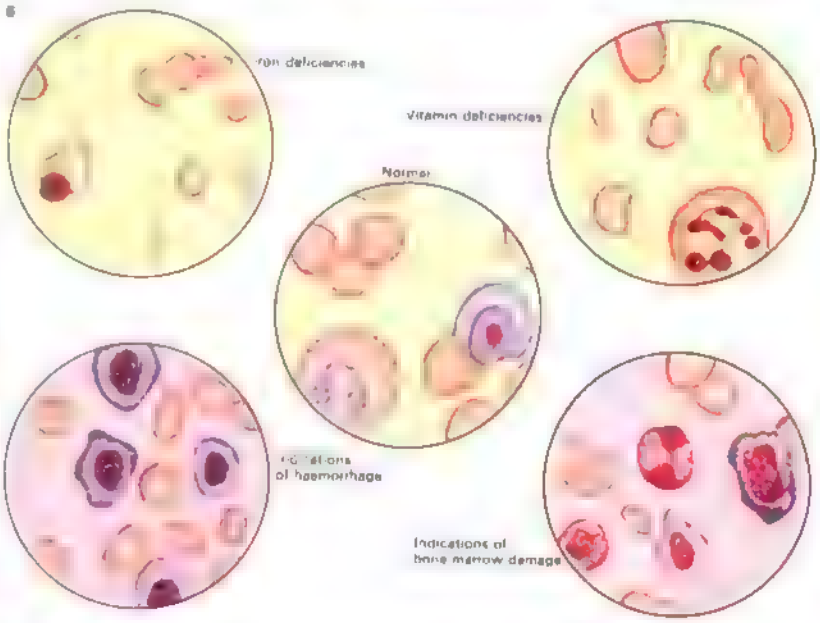
Blood plasma contains various factors needed for the normal clotting mechanism. Deficiency of one, the anti-haemophilic factor, leads to haemophilia [6]. Deficiency in another causes the much rarer Christmas disease which has similar effects. Treatment with anti-haemophilic factor is needed as soon as possible after the start of a spontaneous bleed to prevent further damage [7]. Unfortunately the factor is in short supply and lasts in the body less than 24 hours. The hope for the future is that sufficient will be made available so that haemophiliacs can treat themselves at home



Diseases of the circulation affect the heart, the blood and the network of arteries, veins and capillaries that carry it around the body

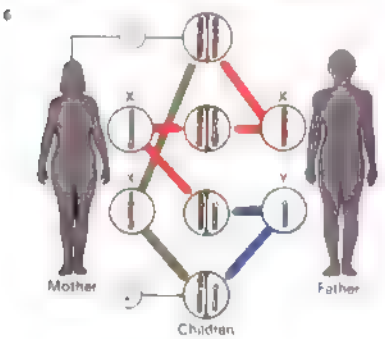
4 Veins normally have a special system of valves that enables the blood to travel towards the heart but not back [A]. People whose occupation involve long periods of standing are prone to varicose veins. This condition occurs when the valves do not function properly, the veins becoming stretched and distorted [B]. Varicose veins of the lower end of the bowel are called piles (haemorrhoids). If the affected veins fail to respond to simple treatment they will sometimes require injections or surgical removal. When varicose veins become infected and inflamed it is a form of phlebitis [C]





6 The gene causing haemophilia (green) is carried on one of the female (or X) sex chromosomes controlling blood clotting. It does not show in a female carrier [1] because her other X chromosome (red) is normal, but if transmitted to the son [2] of a normal male, his Y chromosome (blue) cannot balance it

7 In haemophiliacs, spontaneous bleeding occurs into the joints, which become hot, swollen and painful. The patient must receive anti-haemophilic factor quickly to stop the bleeding because it can lead to a crippling arthritis. The use of ice-packs and analgesics helps to relieve the pain



8 Treatment for leukaemia is aimed at preventing the abnormal white cells from reproducing themselves. Drug treatment interferes with the doubling of the genetic material (DNA) prior to division [A, B] or with the division process itself [C]. Steroid drugs also interfere with division [D]. The theory that leukaemia is caused by a virus is still only a tentative one but if proved, could mean new and revolutionary treatment by the use of vaccines [E]

5 Anaemia produces recognizable changes in the red cells of a blood film. Compared with normal blood, iron deficiency shows up in small, pale cells with reduced haemoglobin content. Iron deficiency may be due to hidden bleeding rather than dietary insufficiency and needs investigating. Lack of either of the vitamins B_{12} or folic acid produces large, pale, abnormally shaped red cells and white cells with multiple nuclei. Diets that are lacking in liver and dairy products like eggs, milk and cheese (containing vitamin B_{12}) or fresh fruit and vegetables (folic acid) can cause these anaemias, but even with an adequate diet the vitamins that are needed may not be properly absorbed. Pernicious anaemia for example, prevents the proper absorption of vitamin B_{12}. After severe haemorrhage or damage, bone marrow may release into the blood red cells that are immature or abnormal in some way

Diseases of the digestive system

Disorders and diseases can occur in any part of the gastrointestinal tract. [illegible] sometimes [illegible] vomiting affects them occasionally but if it becomes frequent and persistent then the chances are that disease has affected the digestive system in the region of the oesophagus, stomach or duodenum or perhaps the pancreas, liver or gall bladder.

The formation of ulcers

The oesophagus passes from the thorax to the stomach through an opening in the diaphragm to which it is loosely attached. From needle age this attachment weakens and part of the stomach can become pulled up in the opening; this hernia is quite common. After a meal the acid contents of the stomach can reflux back into the oesophagus causing pain and eventual ulceration. The pain associated is posterior to other and heartburn and is one of the symptoms behind the medical adage "if a patient complains of his heart, examine his stomach, if he complains of his stomach, examine his heart". The reflux may cause regurgitation of fluid into the mouth, probably after bending or straining. Occasionally surgery may be needed but the usual treatment consists of a weight-reducing diet plus advice on sleeping in an upright position.

The term 'peptic ulcer' usually refers to an ulcer found at the lower end of the oesophagus, in the stomach or in the duodenum [1]. The cause of ulcers is thought to be increased secretion by the stomach of acid and the enzyme pepsin. This acid pepsin mixture secretes strong enough to overcome the natural protection of mucus in the stomach and eats away the digestive tract lining. Alcohol, smoking, stress and O group blood type are all associated with ulcers.

The pain that an ulcer causes is usually felt centrally in the upper part of the abdomen [5]. Its occurrence is related to meal times, but other factors such as between [illegible] taking food or antacids usually relieves the pain, as does vomiting.

The medical treatment of all ulcers involves relief stress with rest in bed, a bland diet, the stopping of smoking and the administration of drugs such as antacids and compounds derived from liquorice root.

There are many disorders in which the main defect stems from the inability of one or more of the organs to absorb the essential items. Malabsorption leads to [illegible] with weight loss, [illegible], diarrhoea and vitamin deficiencies.

Malabsorption disorders

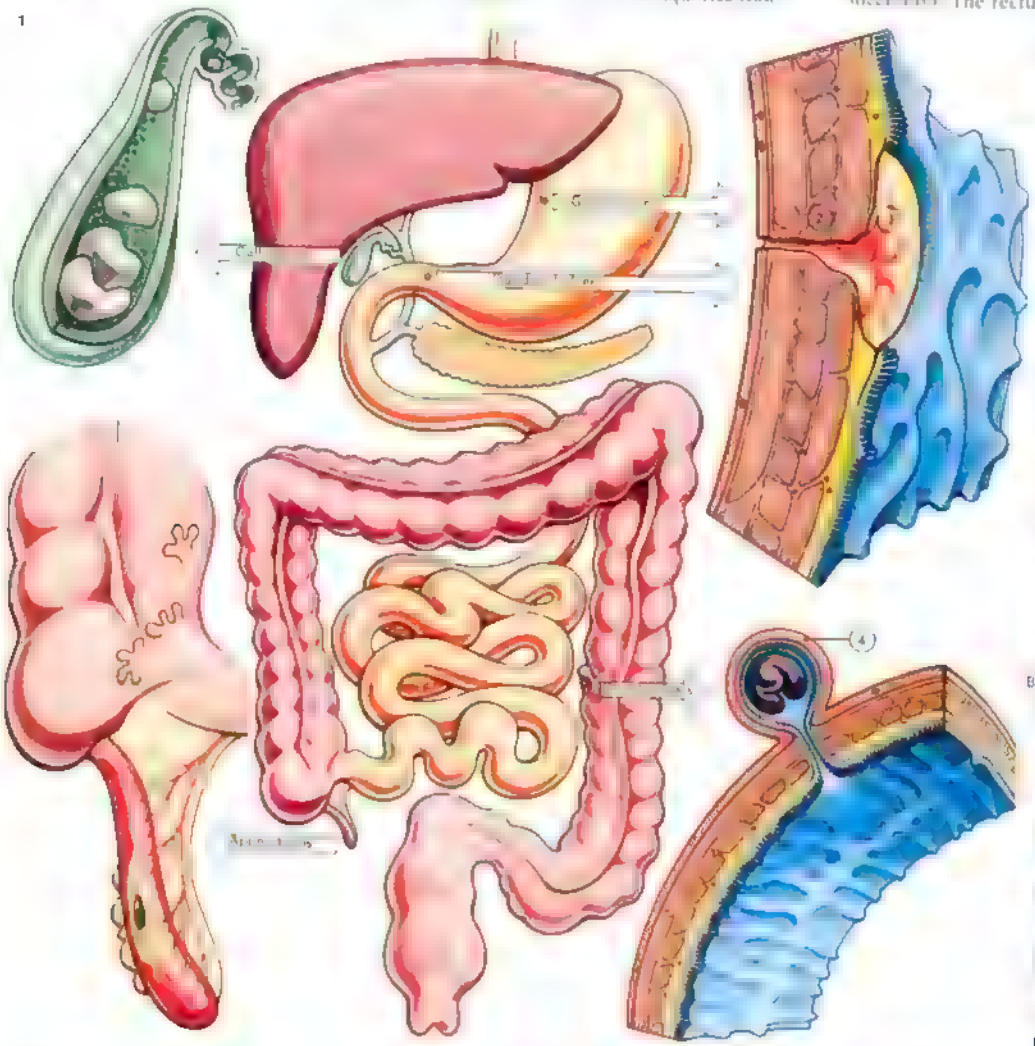
Two of the most important malabsorption defects are coeliac disease and Crohn's disease. Coeliac disease is due to an inability to cope with an allergy to gluten, a protein found in wheat and other cereals, while Crohn's disease is a inflammation of the end part of the small intestine (the terminal ileum) and is also known as regional ileitis. Crohn's disease can sometimes affect other parts of the small bowel and colon. Typical symptoms include abdominal pain, diarrhoea, loss of weight, [illegible] and anaemia. Medical treatment includes bed rest and a diet low in roughage.

Ulcerative colitis is an important disease affecting the large bowel. Its cause is a mystery but it involves inflammation that leads to ulceration. The rectum alone or sometime

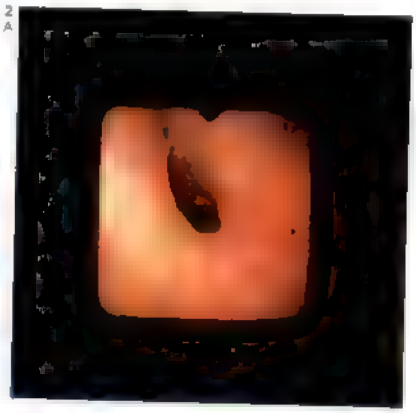
CONNECTIONS

See also

1

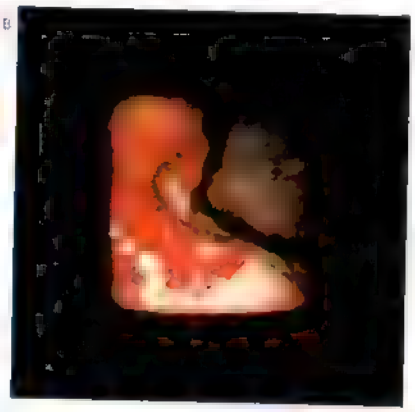


2 A



2 A fibre optic endoscope [illegible]

B



1 Five common disorders of [illegible] stomach or duodenum may penetrate the [illegible]

[illegible] fifty [illegible] **fat forty female and fecund.** They are [illegible]

[illegible] **Appendicitis can also** demand surgery [illegible]

[illegible] **diverticulitis** [illegible] caused by weakening of the bowel wall and in [illegible]

3 Gall stones are most [illegible] **cholesterol and the bile pigment bilirubin.** Stones can be [illegible] of cholesterol [illegible]

3

the whole colon, may be affected. Painful attacks of blood-stained diarrhoea are typical and in severe cases high fever, anaemia, exhaustion and collapse can occur.

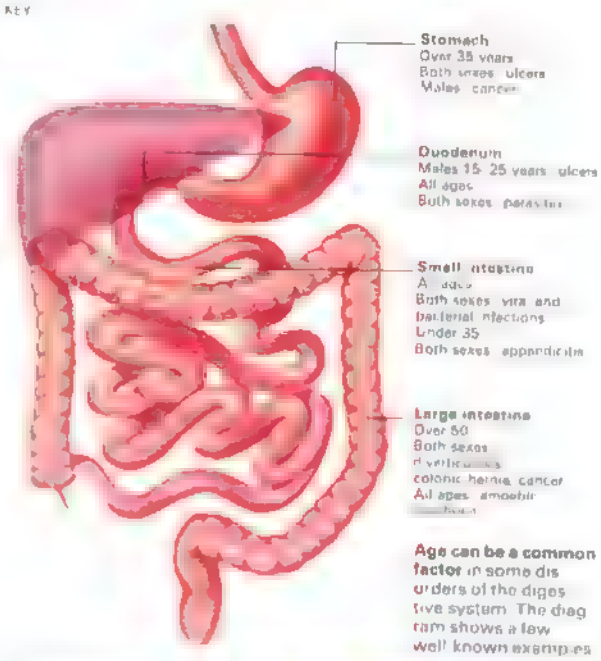
Cancer of the intestine, especially of the stomach and colon, is second only in importance to cancer of the lung. Symptoms include non-specific pain, a change of bowel habits, loss of weight and intestinal bleeding. In cancer of the stomach loss of appetite, nausea and vomiting are usual and in cancer of the bowel blockage by faeces – intestinal obstruction – may take place. After diagnosis surgery is the usual treatment.

Jaundice diseases

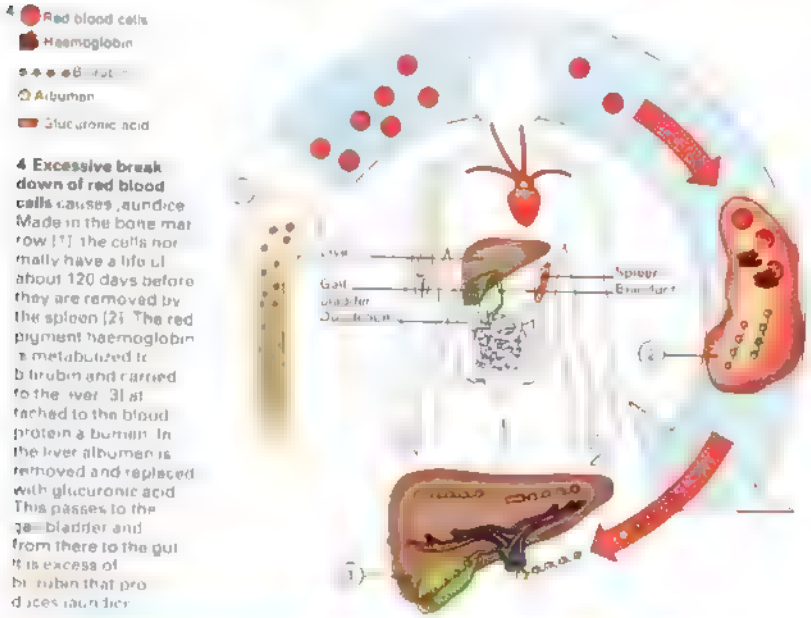
The liver and gall bladder, organs vital to the orderly working of the digestive system, produce jaundice when they malfunction. Jaundice is a yellow discoloration, most noticeable in the skin, that arises when the amount of pigment bilirubin in the blood is above normal [4]. Liver disease or the presence of gall stones [1, 3] in the gall bladder will cause this. The liver diseases hepatitis and cirrhosis are well known for causing jaundice. Cirrhosis, associated with excess alcohol intake, involves the replacement of normal liver cells with fibrous tissue.

Hepatitis is of two types and both are caused by different viruses. Infectious hepatitis has a relatively short incubation period (15–35 days) and like many infectious diseases is spread by droplets (as in a sneeze). Serum hepatitis has a longer incubation period of 40–180 days and is passed in blood transfusion. Both diseases cause inflammation of the liver cells and this prevents them from functioning normally and causes an excess of bilirubin which in turn causes the typical jaundiced coloration of the skin [4].

Apart from the hepatitis viruses, other micro-organisms can affect the digestive tract. Most virulent of these are the *Salmonella* or "food-poisoning" bacteria, which are also responsible for typhoid [6]. Bacteria also cause the inflammation of the appendix and thus appendicitis [1]. Other invaders of the digestive tract are amoebae and parasites such as roundworms and tapeworms. All these organisms can cause severe loss of weight and debilitation.

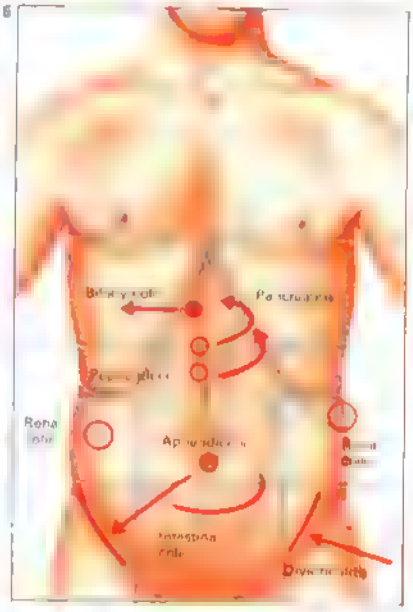


Age can be a common factor in some disorders of the digestive system. The diagram shows a few well known examples.

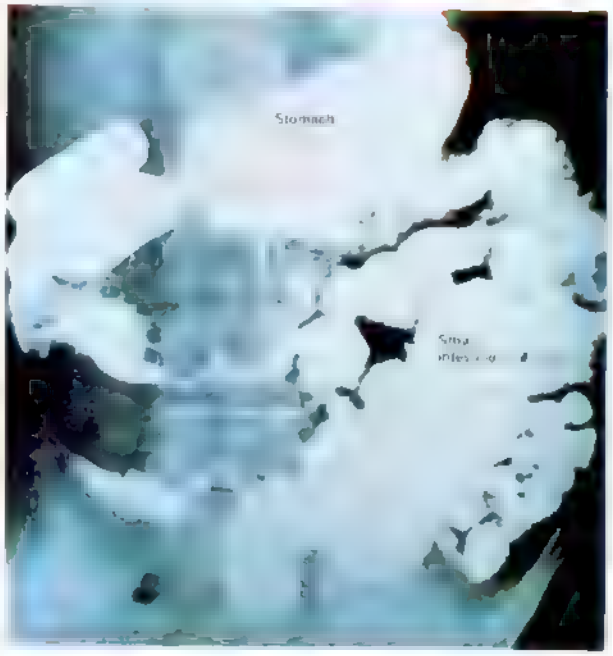


4 Excessive breakdown of red blood cells causes jaundice. Made in the bone marrow [1] the cells normally have a life of about 120 days before they are removed by the spleen [2]. The red pigment haemoglobin is metabolized to bilirubin and carried to the liver [3] attached to the blood protein albumin. In the liver albumen is removed and replaced with glucuronic acid. This passes to the gall bladder and from there to the gut. It is excess of bilirubin that produces jaundice.

5 The acute abdomen is the name given to a group of disorders associated with abdominal pains. Some of them are shown here. The colics are pains that wax and wane in intensity. Intestinal colic can occur in food poisoning or intestinal obstruction. Biliary colic is not true colic but an intense, constant pain starting in the middle and moving to the liver. The pain from a perforated peptic ulcer or inflammation of the pancreas is persistent and moves through to the back. Appendicitis usually starts with a pain in the region of the umbilicus which moves to the area of the right groin. The pain of diverticulitis is in the left lower abdomen.



6 The bacterium *Salmonella* causes intestinal infections that may vary from mild gastroenteritis to severe typhoid. Bacteria enter the body [A] from infected food or water and incubate for up to 14 days. In typhoid [B] disease afflicts many body parts. After incubation [C] diarrhoea ensues. Other features of typhoid [D] are nose bleeds and rose spots. Bacteria remaining after an attack [E] make a person a carrier.



7 A barium meal X ray is a common diagnostic procedure. The patient swallows a barium-containing mixture opaque to X ray. Conditions that may be revealed include ulcers, various cancers and hiatus hernia of the stomach into the oesophagus.

Diseases of the skeleton and muscles

The skeleton is the framework on which the body is hung and supported, and the muscles provide movement. Considering the extent of movement in life, and the weight that is supported, it is not surprising that disorders of the musculoskeletal system occur. However, not all disorders are the result of wear and tear; inherited disorders may also occur, others may arise spontaneously during life

Wear and tear in muscles

The wear and tear disorders are among the most common Everyone suffers muscular aches and pains at intervals and severe and unaccustomed exercise usually leads to discomfort and stiffness in the muscles the next day. In due course such exercise of muscle will lead to the development of more powerful muscles In the untrained, sudden movements will more commonly lead to muscle strains, tears of parts of the muscle mass, or tears in the tendons which attach muscles to bones. Such tears are often accompanied by bruising in the muscle, and healing can take two or three weeks

Lumbago, or pain in the lower back, may follow muscle strain which results from unaccustomed activity such as digging When back pain is accompanied by pain going down the back of one leg the condition is known as sciatica Here the pain is due to pressure on one of the nerve roots that leave the spinal cord and supply the leg via the sciatic nerve Treatment involves rest in bed on a firm mattress supported by wooden boards. The strain may be on one of the ligaments that run up the spinal column and support the vertebrae

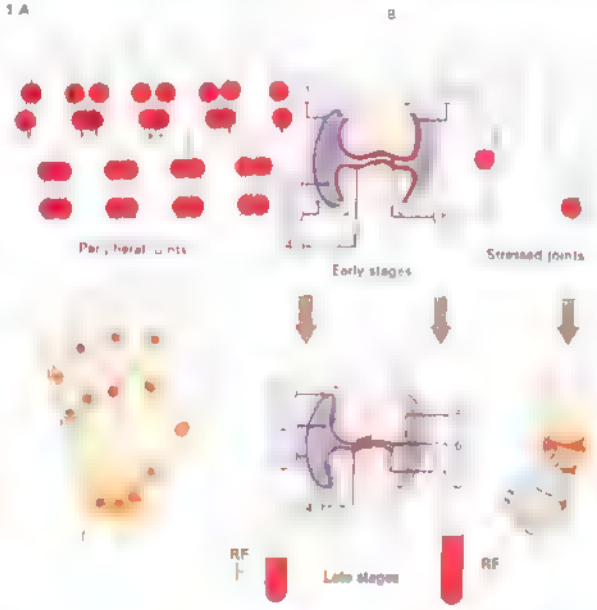
A weakness in the ligaments can allow the disc or pad which cushions each vertebra from the next to slip from its normal site and press upon the nerve root [3]. Treatment of a slipped disc requires rest while the tear heals and the disc returns to its normal place If rest does not achieve this, surgical removal of the protruding portion of the disc may be necessary to alleviate symptoms

Other sites at which injury can occur are in the joints between the bones. Bone ends are covered in cartilage and held together by strong fibrous capsules. The capsule of a joint may be torn, and the cartilage may be damaged In the knee, specialized half-moons of cartilage act as cushions between the femur (thigh bone) and the tibia (shin bone) These two cartilages in each knee are attached to the edge of the wearing surface of the knee joint. Sudden twisting movements of the knee may be enough to dislodge a cartilage, tearing its attachment to the bone. This is painful, prevents full use of the knee, and if it does not settle the offending cartilage may have to be removed. Cartilage injuries are common in sports [6] in which the weight-bearing joints are frequently twisted

Arthritis and its treatment

Over the years the degree of wear and tear becomes excessive and the consequence may be the premature development of osteoarthritis, also known as osteoarthrosis [1], in the joints. This type of arthritis develops in many elderly people in the ankle, knee and hip The results are that the joint narrows, movement is restricted and pain and deformity occur Osteoarthritis in the hip can be the most disabling, but it is now possible to replace the hip joint with metal and plastic substitutes and thus restore full mobility

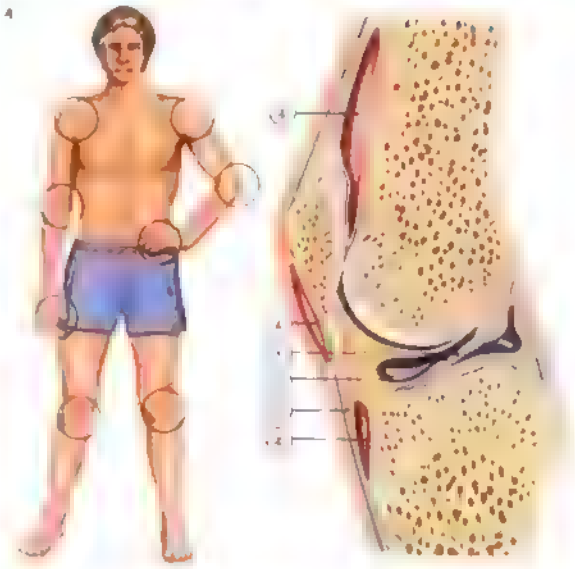
CONNECTIONS



1 Arthritis means inflammation of the joints. In rheumatoid arthritis (A) the synovial membrane [1] becomes inflamed and thickened and produces increased synovial fluid within the joint [2] The capsule and surrounding tissues [3] become inflamed, while joint cartilage is damaged [4] Peripheral joints, as in the feet and hands, are involved Blood tests reveal the presence of rheumatoid factor [RF] and the rate of red-cell sedimentation in a test tube rises. Osteoarthritis [B], a degenerative disease, involves thinning of cartilage [5] loss of joint space [6] and bone damage [7] Heavily used or weight-bearing joints are affected Blood tests are normal

2 The hands are common sites of rheumatoid arthritis. An X-ray shows the typical deformity of the joints and the deviation of the fingers towards the little finger side The causes of rheumatoid arthritis (not clearly defined) may include infection hereditary factors, emotional disturbances and physical overuse of the affected part Treatment may involve the administration of anti inflammatory drugs

3 The spinal column (A) is a series of bones (B) surrounding the spinal cord [1] and separated by discs or cushion like pads [2]. These may slip out of their normal place if the surrounding ligaments tear due to strain or degenerative changes. The disc between fourth and fifth lumbar vertebrae [3] here protrudes towards the back of the vertebrae Pain, spasm of the muscles and restricted movement are common results. The fifth lumbar vertebra [C] (from above) shows the nerve roots [4] leaving the spinal cord The position of the normal disc [5] is shown (pale), while the herniated disc [6] is seen pressing on the nerve root



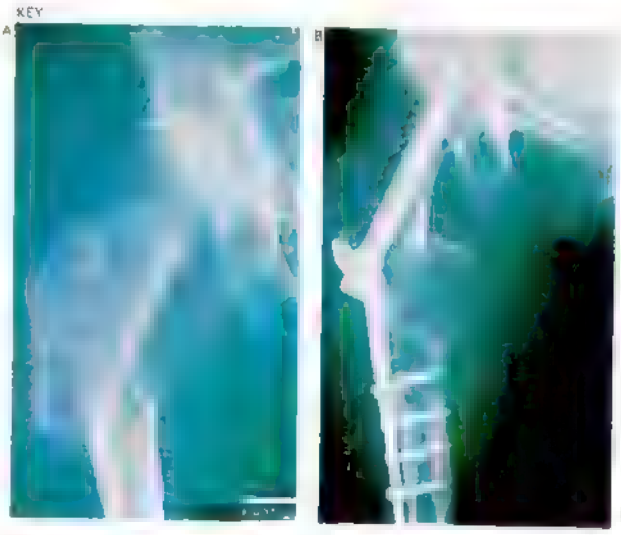
4 Joints are well lubricated by synovial fluid [1] produced by a synovial membrane [2] Around many joints pass tendons [3] which are protected from rubbing and excess friction by strategically sited sacs or bursae [4] which are similarly lined and lubricated Local damage overuse or pressure can lead to inflammation and excess production of lubricating fluid (bursitis) Water on the knee, or housemaid's knee, is typical Trigger finger and frozen shoulder are often due to synovitis, while tennis elbow results from damage at the attachment of a tendon To avoid further strain the joint should not be overused but exercised gently, immobilizing it can lead to stiffness and may be harmful

The other common arthritis, known as rheumatoid arthritis, is not a result of overuse. Women are more likely to be affected than men. The small joints of the hands and feet become damaged and deformed, and elbows, shoulders and ankles are also affected. The disorder is sometimes part of a more generalized body disturbance, and evidence of widespread inflammation can be detected in blood tests. In both types of arthritis, symptomatic treatment for the pain with analgesics is beneficial, but suppression of the inflammation may be required in rheumatoid disease. This is often achieved by administering other drugs.

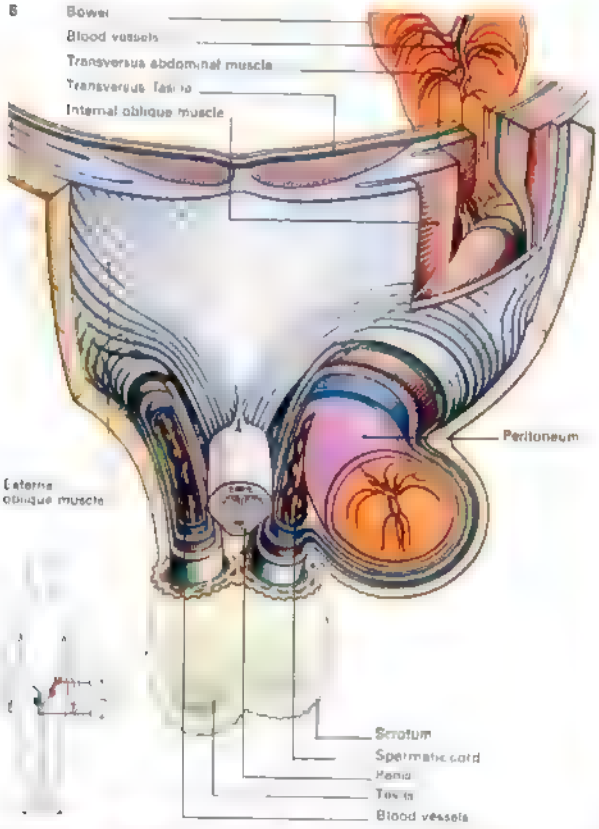
Some musculoskeletal disorders are congenital, that is, present from birth, for example, congenital dislocation of the hip is sometimes found in infants [7]. The hip joint does not form fully for many years after birth, and initially the socket in which the ball of the femur is held is very shallow. It is quite easy to dislocate the head of the femur from the socket. In a few children this occurs spontaneously and, if not corrected, will cause maldevelopment of the hip. Treatment is simply to splint the legs so that the joint is held correctly and cannot dislocate. Although ungainly, such treatment at an early stage is effective. The longer that diagnosis is delayed the more difficult is the cure and babies are examined shortly after birth for evidence of this uncommon condition.

Treatment by immobilization

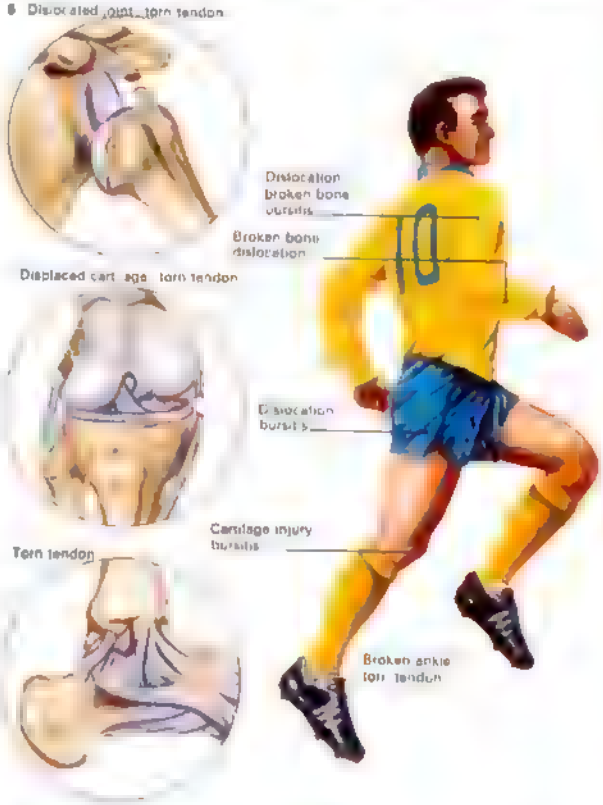
Immobilization of an injured part of the musculoskeletal system can be required for many different troubles, such as fractures of bones [Key]. Plaster of paris casts may be adequate if the parts of the bones are well aligned. Otherwise internal fixation can be undertaken, in which the bones are held together with pins or plates screwed into the fragments. It is important at every stage after the operation, including the period during which the limb is splinted, that the patient is made to exercise as much as possible, because during the enforced rest much wasting of the unused muscles will have occurred. This can considerably slow a full recovery, for building up muscles always takes time, effort and dedication.



Fractures of the neck of the femur [A] are common in old people, who have quite brittle bones. The X-ray shows how pin and plate surgery [B] to the broken fragments can help to speed healing and recovery. The bones of young people break less easily than those of old people and when they do break they often cause less damage to surrounding tissue.

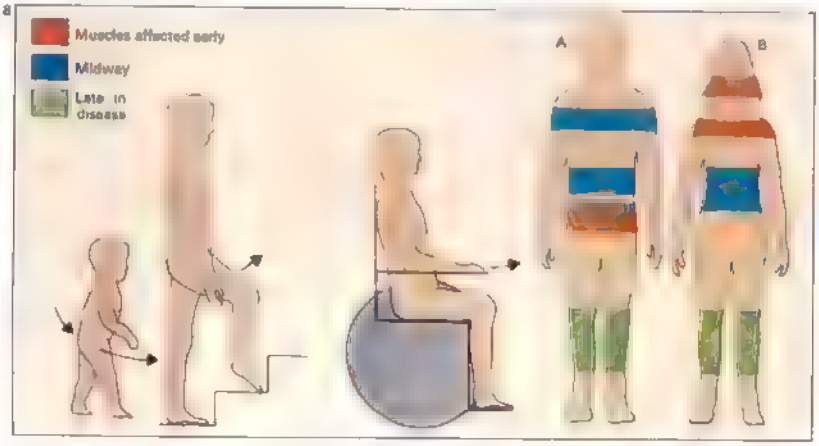


5 A hernia, often called a rupture, is the passage of any organ into an abnormal site. Common are those where the bowel [1] slides into the groin [2]. A weakness develops in the abdominal muscles. A loop of bowel insinuates itself along the path of the spermatic cord in the male and appears as a bulge in the groin. Such an inguinal hernia may be controlled by a truss or can be treated surgically.



6 In sports such as football, undue stress may be put on bone, tendon and muscle, and physical fitness is essential to minimize the possible risks. Shown here are the different kinds of damage that the various parts of the body suffer from. Sudden movements and collisions result in dislocated joints or broken bones in the shoulder, the arm and the leg. Cartilages in the joints may be displaced, while muscles and tendons are strained or torn.

7 Congenital dislocation of the hip is a condition found at, or shortly after, birth. It is due to an increased ability of the head of the femur to slip out of the developing socket in the pelvic bone. In this X-ray, the joint on the left is normal, the one on the right displaced. Characteristic clicking of the hip is checked for in the weeks after birth.



8 Muscular dystrophies are inherited degenerative diseases that cannot be cured. The Duchenne type [A] affects the pelvis, shoulders, trunk and later the limbs. The disease is first noticed as the child learns to walk with a characteristic waddling gait. He falls frequently and has difficulty in climbing stairs and rising from a sitting or lying position. A wheelchair becomes a necessity. The facio scapulo humeral type [B] affects the face, shoulders, trunk and finally the legs.

Diseases of the nervous system

The nervous system is made up of the central nervous system, comprising the brain and spinal cord, and the peripheral nervous system that carries messages to and from all parts of the body. These two parts of the nervous system vary greatly in their reaction to damage and disease, for while the cells of the peripheral nervous system are resistant to wear and tear, and have some ability to regenerate if partially destroyed, the cells of the central nervous system are susceptible to damage and once lost are irreplaceable.

Physical damage to the brain

The brain is often compared to a computer, but unlike a computer does not totally break down from minor damage. The skull provides fairly good protection but a knockout blow on the head may cause temporary amnesia (loss of memory) or concussion, which can be thought of as temporary brain damage. Even so, recovery is the rule.

Fracture of the skull [1] can cause physical damage to the brain. This, in turn, can result in cessation of activity in body muscles whose actions are controlled directly by the brain. Because the skull in an infant is not fully closed the brain is more liable to damage, and physical damage, although rare, can occur during birth.

The more usual cause of damage is from within. In some people the arteries of the brain are fragile and can bulge out like weak tyres, a condition doctors call aneurysm. The great danger of this is that the artery may burst, causing bleeding into the brain. This is a cerebral haemorrhage in which the patient may become paralysed, and which can be fatal. A stroke, which involves a sudden loss of cerebral function, may be caused by the haemorrhage of a central vessel or by a blood clot or thrombus within a vessel.

Tumours in the brain cause physical damage by pressure they exert and other damage by the loss of brain tissue to the tumour. With modern techniques such tumours can in some cases be easily located and removed by surgery [2].

Effects of toxic chemicals

The brain is also sensitive to chemical damage. This can arise in various ways. Some of the chemicals that find their way into the diet can affect the brain. Chronic lead poisoning, for example, causing tumours and paralysis, was once a problem in soft-water areas when lead piping was used for domestic water supplies, because soft water dissolves lead. In some disorders of metabolism normal body chemicals build up to such a level that they cause damage to the brain. Phenylalanine, a constituent of body proteins, is one such substance. In the inherited disease called phenylketonuria it is created to excess and causes mental retardation.

The chemicals harmful to the nervous system (neurotoxins) can be produced by bacteria that invade any part of the body and reach the brain via the bloodstream or via nerve trunks. The neuritis of diphtheria is caused solely by the toxin the diphtheria bacilli produce. Some bacteria release toxins that affect only nervous tissue [3]. The food poisoning organism *Clostridium botulinum* produces a neurotoxin that causes fits, double vision, paralysis and death.

Some other micro-organisms seem to infect the central nervous system preferen

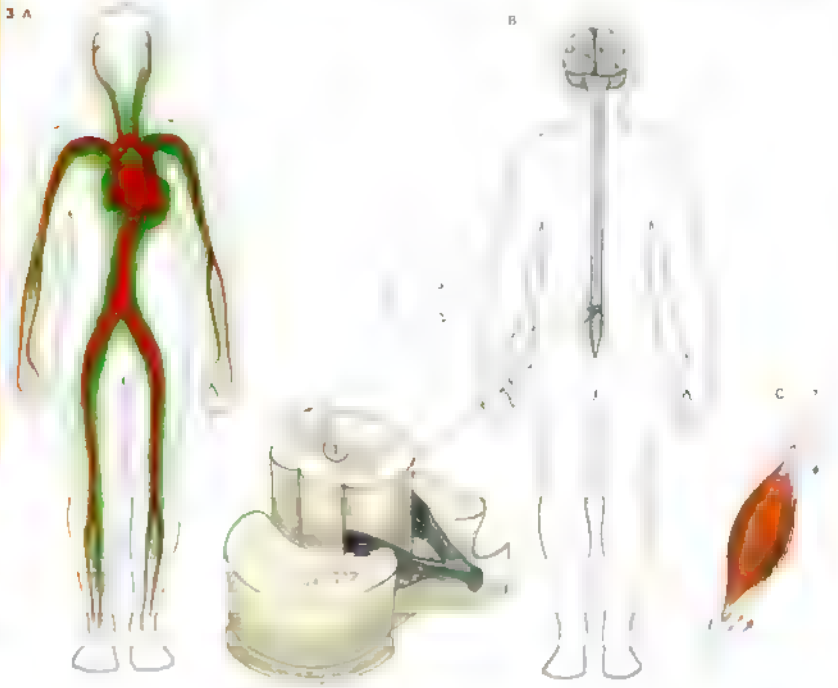
CONNECTIONS

See also

1 Road accidents are a significant cause of brain damage and [illegible] in a series of experimental [illegible] knock down a child dummy [A] but scoop up an adult dummy [B] [illegible] The [illegible] bonnet picks up the child dummy and reduces head impact. It [illegible]

2 Brain surgery is most often performed after injury to the brain associated with a skull fracture. Such fractures are most commonly caused by car accidents. Brain surgery is usually directed towards stopping bleeding [illegible] the removal of blood clots (haematomas). Other types of brain [illegible] that necessitate surgery [illegible] and the rupture of weak spots or aneurysms in the walls of blood vessels in the brain. Diagnosis of brain damage and disease has been [illegible] since the 1970s by the EMI [illegible] machine the brain can be visualized in serial "cross-sections"

3 Tetanus, popularly called lockjaw, is caused by wounds infected with *Clostridium tetani* bacteria. These produce a poison that does not affect the wound but is carried [A], often by the bloodstream, to the central nervous system [B] where it becomes concentrated in the nerve cells of the spinal cord [1]. The normal control of muscle action that takes place in the spinal cord is completed by messages from the brain that modify signals passing out to the muscles via the motor nerves. The tetanus toxin acts by blocking the biochemical reactions essential to the completion of the nerve pathway. Thus the muscles cannot relax and go into spasm [C]



tially. The rabies virus, for example, specifically attacks the nervous system [7], as does the virus causing poliomyelitis [5] which, by attacking groups of motor nerves, paralyses the muscles whose movements they control.

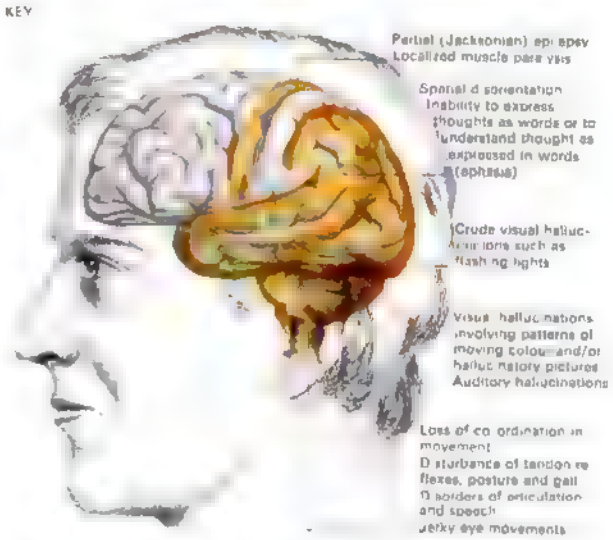
Some micro-organisms that normally cause infections elsewhere find their way into the brain tissue (causing encephalitis) or its surrounding membranes or meninges (causing meningitis). Untreated syphilis can eventually cause an encephalitis, as can measles in about one case in 3,000. Pneumonia and tuberculosis can both cause meningitis, but as with all microbial infections the modern advances in vaccination [4] have reduced the incidence of disease, and antibiotics have in some cases controlled the spread of infection.

In degenerative diseases of the nervous system the nerve tissue undergoes chemical changes that stop it working properly. In multiple sclerosis, the most important of these, there is patchy hardening of the myelin that insulates the nerves, causing an intermittent loss of function whose locality depends on the exact bundles of nerves affected. Muscle paralysis and speech disorders are two examples of degenerative diseases.

Symptoms of disease

Because some parts of the brain perform distinct jobs the symptoms of disease are often dictated by the area affected. Aphasia, for example, a disturbance in the ability to speak, write and comprehend words is due to damage in the area that controls speech. Parkinson's disease, in which the body trembles when at rest but becomes rigid at other times, is due to destruction of nerve cells at the base of the brain that are involved in the control of movement. The new drug L-DOPA now helps 40 per cent of all patients with Parkinson's disease. Yet one of the most remarkable things about the brain is that damage to some areas produces no detectable defect at all, as if the other cells take over the activities of the damaged ones.

Physiologists specializing in the study of the nervous system have made great advances in discovering the causes of disease but epilepsy [8] remains a mystery, although scar tissue, a tumour or fever can be causes.

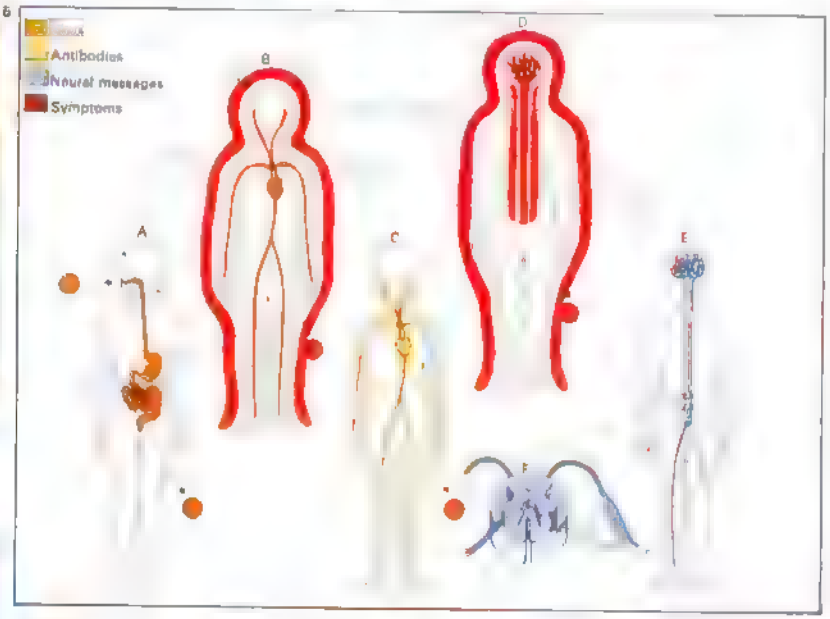


The symptoms that result from brain damage may reflect the particular function of the area in question. Some examples of such symptoms are indicated here. The specific effects of certain brain injuries have been of great use in producing physiological maps of the brain.

4

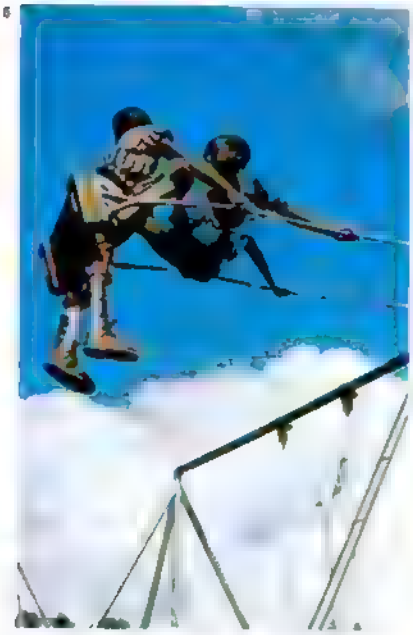
4 A vaccination campaign against polio started in Europe and the USA in the 1950s has almost obliterated the disease from those regions. Today babies receive polio vaccine in their first few months of life.

6



5 The polio virus enters the body by the nose or mouth. A incubates in the intestine and then enters the bloodstream (B) causing headaches, fever and vomiting. The body now 'fights back' by releasing antibodies (C), which may effect a complete recovery and also confer immunity against another attack. Or the virus may continue to multiply, reach the brain and spinal cord and destroy nerve tissue (D). The result is that the nerves can no longer send out messages to the muscles (E) to effect contraction and the muscles (usually of the limbs) become paralysed. The enlarged section of the spinal cord (F) shows the damaged area (black) and its 'dead' nerve

5

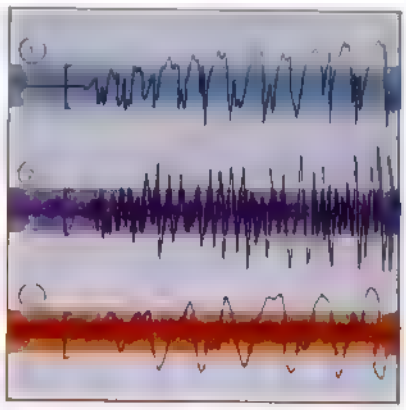


6 At a rehabilitation centre in Rwanda victims of polio – a disease also known as infantile paralysis because of its preferential attack on children – receive treatment to encourage paralysed muscles back to action.

7

7 Rabies is a fatal disease caused by a virus that selectively attacks the brain. It is caught from the bites of infected animals, particularly dogs. Louis Pasteur (1822–95) developed the first rabies vaccine from the saliva of 'mad' dogs.

8



8 Epilepsy is not so much a disease as a symptom of a sudden disturbance in the brain's normal function. During an epileptic seizure the brain shows characteristic electrical rhythms that can be recorded on an electroencephalograph. During a "grand mal" (1) the patient usually loses consciousness and will fall unless supported. Most seizures last a few minutes. In a "petit mal" (2) the unconsciousness may be momentary. Focal epilepsy (3) is characterized by hallucinations.

Diseases of the skin

The skin, the body's covering organ, has a triple susceptibility to disease [Key]. First, because of its location, secondly because of its dependence on the rest of the body; and thirdly because the skin itself is prone to faulty functioning.

Attack from fungi

Infestations of the outer skin layer (the epidermis) are usually caused by microscopic fungi that use dead cells and skin secretions as food. These fungi are particularly prone to grow in humid protected regions, the armpits, groin, hair and feet in particular. Different types of fungi [5] grow in the different regions, and although the generic term ringworm [6] may be applied to them, they have nothing to do with worms. The mouth and vagina are also liable to colonization by the fungus *Candida albicans*, or thrush [7]. These fungal infestations are not usually harmful although they cause intense irritation. They must be treated, however, as severe complications can result

The skin is also colonized by billions of bacteria. In general these do no harm; indeed, because their prescence makes it difficult for some dangerous bacteria to get a hold, they in fact do good. These beneficial bacteria use organic substances in sweat as food and the major disadvantage of their presence is that the chemical transformation of sweat they cause gives rise to the unpleasant odours of stale sweat. Other less wholesome bacteria can be present on the skin and if these manage to penetrate it through a cut or crack, they can infect a wound or may form an abscess. Similar bacteria result in the formation of spots or, more seriously acne [1], in which a hair follicle that has been blocked by excess secretions becomes infected. Impetigo is another bacterial skin infection. It causes itching and is contagious.

Itching and burns

In the UK insect infestations of the skin are now rare, but less than 50 years ago infestation by mites, lice and fleas were common Itching caused by insects was so common that infestation by the scabies mite, burrowing into the skin, was called "the itch"

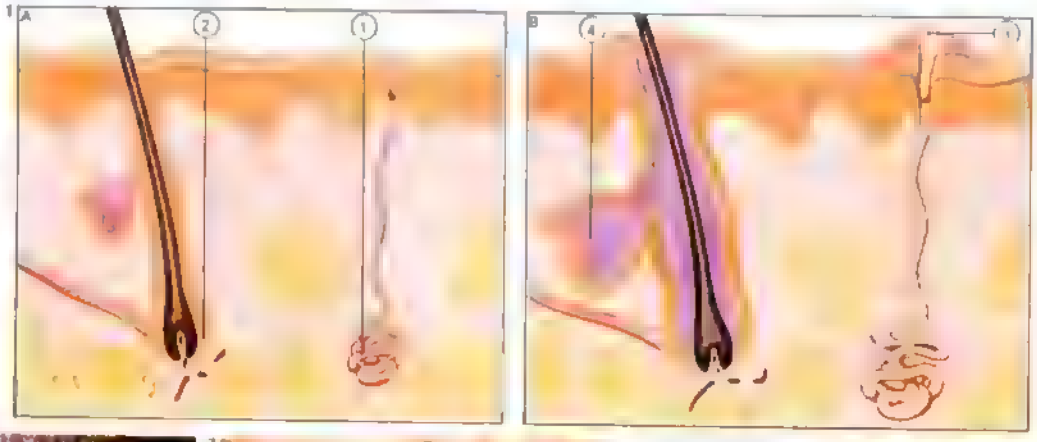
Itching is a non-specific response to any irritation of the skin. It can be psychological in origin, or be caused by a sudden fall in temperature or by excessive sweating. Where redness and swelling are present, they are likely to represent a local toxic reaction to plant or animal stings, for example, or an allergy [4] to a chemical. Eczema and dermatitis are both skin diseases often caused by allergic body responses. They may also be due to the irritant effect of chemicals that cause direct damage. Many chemicals, ranging from epoxy resins to washing powders, can be responsible

Burns usually result from exposure of the skin to high energy (heat or radiation), although they may be caused by acids. They are divided into three degrees depending on their severity and the depths of the burn First-degree burns involving redness of the skin occur in sunburn [11] and mild overexposure to other sorts of radiation such as X-rays or gamma-rays. Second-degree burns are more severe but still heal without a scar, while third-degree burns involve the peeling of all the epidermis and consequent painful exposure of the dermis and its nerve

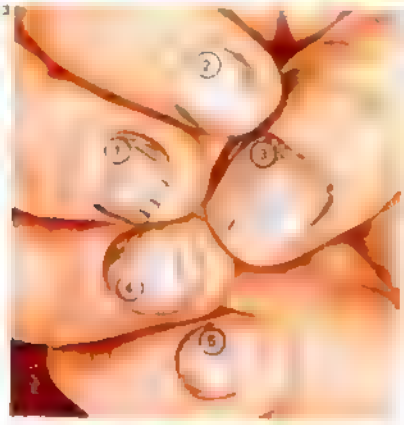
CONNECTIONS

See also
Skin and hair
Allergy

1 **The sweat glands** [1] and hair follicles [2] of normal skin [A] are liable to disorders and infections [B] In hot, humid climates, the openings [3] of the sweat glands may become blocked, resulting in rupture of the gland and the inflammation known as "prickly heat" Bacterial infection of the oil-secreting gland [4] that is attached to each hair follicle is one of the causes of acne, in which pustular bodies develop

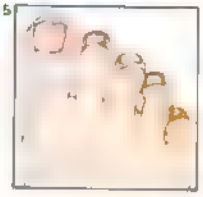


4 **Nettle rash or urticaria** is a reaction shown on the skin in response to the absorption into the bloodstream of a substance to which the body is allergic The substance is said to act as an antigen in nettle rash, also called hives, the itching is caused by the release of histamine, while body enzymes produce the red rash Strawberries, eggs, nuts and shellfish are other common allergy producing agents

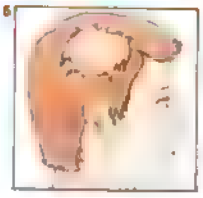


3 **The nails**, which develop from the skin beneath them called the matrix, may reflect deficiencies in the diet, for lack of calcium can make them brittle [1], or flecked [2] A congenital abnormality of the nail producing tissue can result in ridged nails [3] Bitten nails often "ingrow" [4] into the surrounding skin, taking with them infective bacteria. The same bacteria gain entry if the cuticle is destroyed [5]

2 **Scars form over wounds and burns.** Corpuscles invading the wound produce fibrous tissue that joins the broken dermis, and this becomes covered with cuticle At first the scar is delicate and the presence of many capillaries gives it a red appearance As the wound heals, the fibrous tissue thickens and the capillaries are lost, so an old scar is hard and white The faster the edges are brought together the less the scar shows

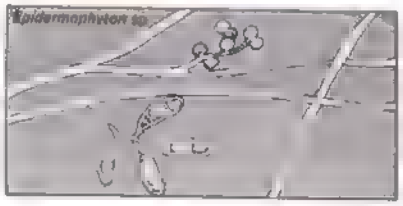


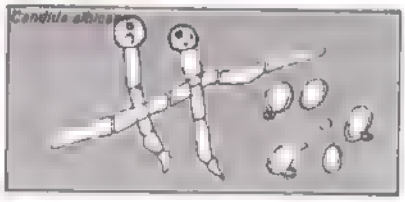
5 **Athlete's foot** (misnamed *Tinea pedis* or ringworm of the foot) is the most common fungal infection of the skin The affected skin becomes sodden and white, peeling off to reveal a red raw area



6 **Ringworm of the scalp**, also a fungal infection, causes lesions of the scalp and loss of hair from the affected areas Common types of ringworm are easily cured and hair regrows

7 **Thrush** is the name given to the white patches of infection on the throat, tongue, lips and palate in young children The same fungus causes an uncomfortable vaginal infestation in women





endings, with formation of scars on healing

Scar formation is the skin's method of repair to any severe damage on its surface. Scars are formed after third-degree burns and as a result of wounds, or due to rupture of the skin as when an abscess bursts. People in some cultures deliberately wound the skin to form scars: Prussian duelling scars, for example, or African tribal scars [2]

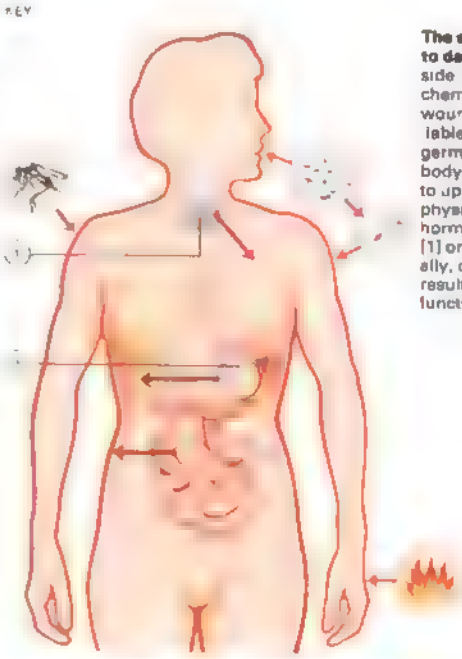
Infective agents

Skin problems by attack from within can be caused by infective agents. The viral diseases of childhood, as well as smallpox [9], are examples of diseases in which the infective agent reaches the skin through the bloodstream. The skin problems – rash [8], itching and the formation of pustules ("pox") – are only one part of the infection and its effects on other organs are the more severe, if less obvious, risks to health; arthritis or osteomyelitis (an infection of bone) can follow.

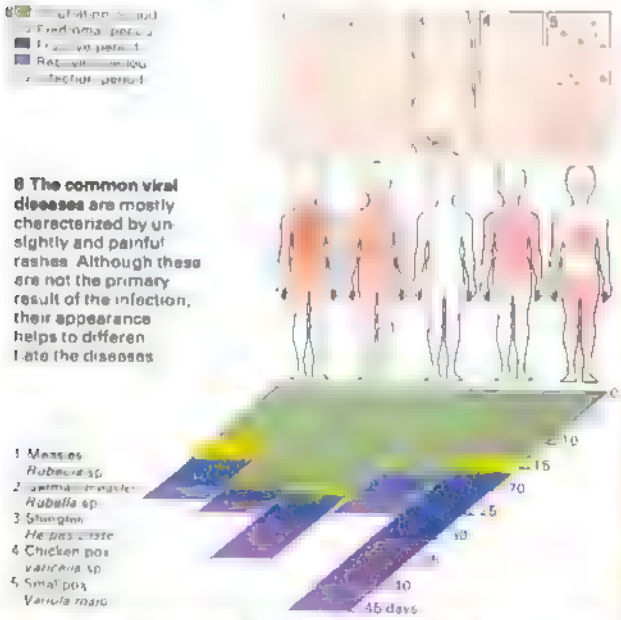
Direct viral attack of the skin is thought to be the cause of *verrucae* (and possibly warts), which seem to be contagious. There are also psychological aspects about them, for many people never get them, despite exposure.

Wrong diet and upset metabolism are probable causes of malfunction in the skin's sebaceous glands, which can become blocked through over-production of oily sebum, producing a local build-up of sebum that can lead to dandruff or acne. Many of the less well-known skin rashes are caused from within the body and affect its metabolism. Red and greasy skin, for instance coarsened in texture (*rosacea*), is a disease apparently resulting from overindulgence and indigestion but it does affect tea drinkers and drunkards alike.

Several skin diseases are associated with a breakdown in normal function of the tissues of the skin itself. The most common of these are baldness, or alopecia, in which the hair follicles apparently wear out, failing to replace hair as it falls, and the whitening of the hair due to loss of pigment. These are both usually senile changes, with a strong hereditary component. Birthmarks, in contrast, are localized abnormalities that involve overpigmentation of the skin (moles) or both overpigmentation and disorders in the blood supply (port-wine stains).

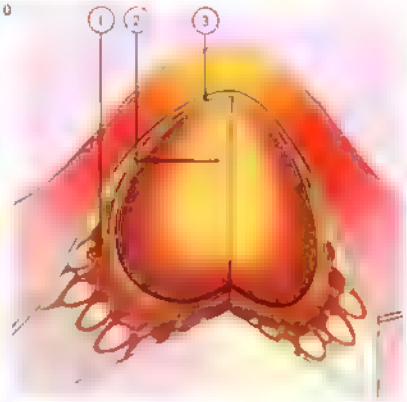


The skin is liable to damage from outside: from parasites, chemicals, burning or wounds. It is also liable to attack by germs that enter the body and is sensitive to upsets in normal physiology caused by hormonal imbalance [1] or disease [2]. Finally, diseases can result from the faulty function of the skin.



8 The common viral diseases are mostly characterized by unsightly and painful rashes. Although these are not the primary result of the infection, their appearance helps to differentiate the diseases.

9 Although the skin rash is superficially similar to chicken pox, smallpox is a far more dangerous disease. By the time the characteristic pus-filled vesicles appear, the disease has already affected the other susceptible organs. The vesicles gradually dry out over the ensuing two weeks when complete recovery occurs. Smallpox is only now endemic in two countries, namely Bangladesh and Ethiopia.



10 When bacteria penetrate the skin there is a local invasion of white blood cells that engulf the invading bacteria and plug the local capillaries [1] to isolate the infection. The mixture of white cells, dead bacteria and tissue fluids [2] build up inside the fibrous capsule [3] that develops round the infected area. If the abscess is not lanced, more pus is produced until the skin ruptures.

11 The most common form of burning is sunburn, caused by over-exposure to the ultraviolet rays of sunlight. Cautious exposure to the sun causes the development of brown pigment which is the body's natural defence against sunlight. Additional benefits include ultraviolet light and also the formation of vitamin D. Sunlight has been implicated, however, as one cause of cancer of the skin.

12 Albinism can occur in any population. It can be total, which is a congenital condition or patchy. The lack of pigment causes sensitivity to strong sunlight. Pink coloration is due to reflection of light from blood vessels.

Diseases of the glands

The glands may be divided into two types according to where their secretions go. Those that release their secretion into a duct or tube are called the exocrine glands [8], while those without such tubes release their product, known as hormones, directly into the blood and are called endocrine glands. The exocrine glands normally release their products to the body surface – for example, sebum to lubricate hair and skin or tears to moisten the eyes – or into the alimentary tract providing the digestive juices. It is alterations in the activities of the endocrine glands that are normally termed endocrine disorders, though glandular fever is a disorder of yet another series of glands – the lymph glands – which are part of the body's defences.

Diabetes: causes and symptoms

The most common endocrine disorder is sugar diabetes (diabetes mellitus), in which insulin lack occurs [8]. The major part of the pancreas is exocrine, producing digestive juices, but about one-tenth of the tissue is direct endocrine, clumps of cells called the islets of Langerhans. Lack of insulin secretion from these islets leads to a failure to control the use of the body fuels and building blocks (sugars, fats and amino acids). In diabetes these are overproduced at the expense of the body tissues and are under-used. Sugars, such as glucose, accumulate in the blood and are poured away as waste in the urine along with excess fluid and salts.

This great loss of water and urine led the Greek Aretaeus, in the first century AD, to call the disease diabetes, after the Greek word for "siphon". Thirst and weight loss occur and high sugar levels encourage infections. The situation can be corrected by insulin injections. When diabetes appears later in life some insulin is produced by the pancreas, but insufficient to keep the metabolism entirely normal. Such patients are often overweight, but with weight loss and controlled intake of carbohydrate food extra insulin is often not necessary to treat their diabetes.

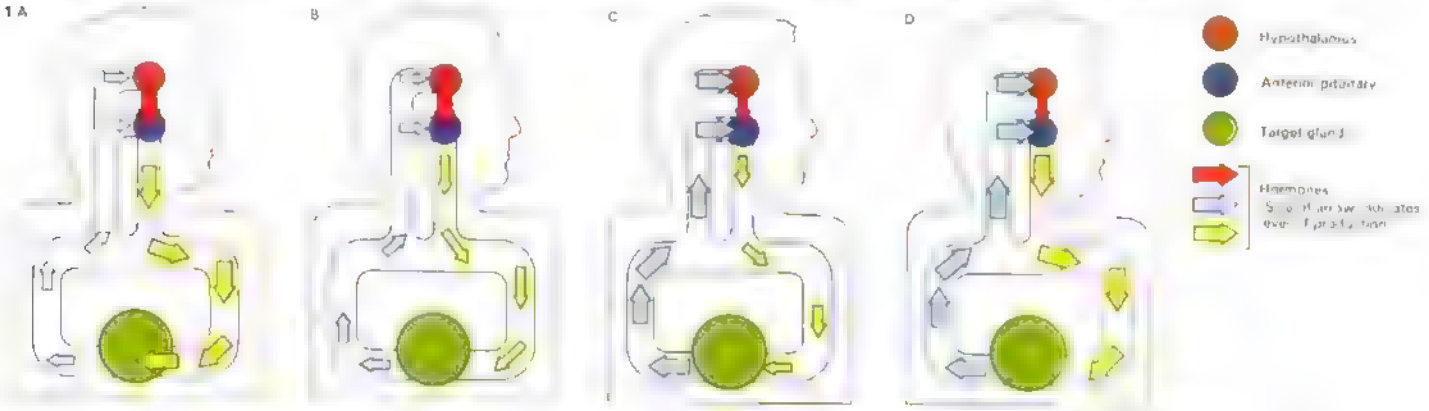
While long-standing diabetes may lead to circulatory disorders, careful control helps to minimize such problems. Diabetes of varying severity may affect two–three per cent of the population in Europe and the United States, but the opposite situation of overproduction of insulin is very rare. It is caused by an insulin-producing tumour of the pancreas, or insulinoma. Blood sugar falls excessively, resulting in hunger, sweating and altered mental ability.

The thyroid and pituitary glands

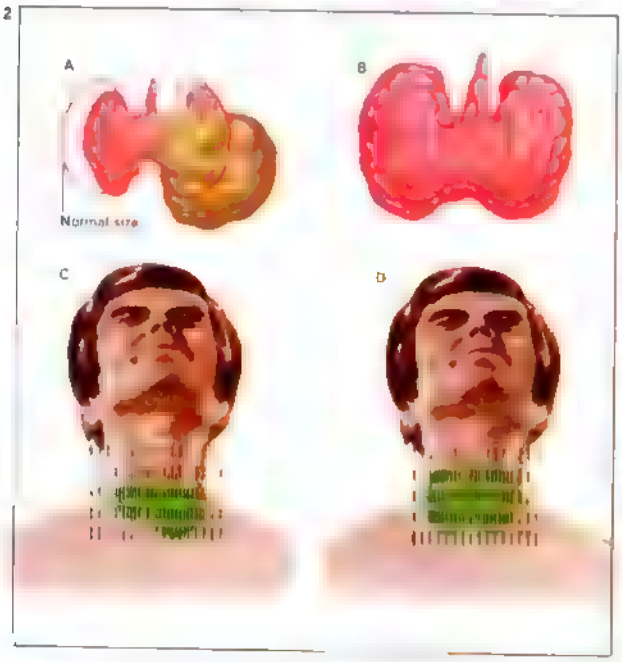
The next most common group of glandular disorders are those affecting the thyroid gland. Overproduction of thyroid hormone leads to increased demands for food and energy by the body tissues, increased heat production, weight loss, nervousness and irritability [2]. It may be treated with drugs or thyroid surgery. A lack of thyroid hormone, from failure of the gland, leads to slowness, apathy, weight increase and susceptibility to cold. It is more common in older people and in those with thyroids that have been overactive in the past. Treatment is by daily thyroid hormone tablets. An enlarged thyroid or goitre is caused by either overactivity or underactivity of the gland. Diets deficient in iodine are one reason for overactivity.

CONNECTIONS

See also



1 If one of the target glands [illegible] which [illegible] pituitary hormone [illegible] hormones [illegible] the thyroid [illegible] Both the hypothalamus [illegible] part of the [illegible] and the [illegible] pituitary [illegible] the base of the [illegible] released, [illegible] the pituitary [illegible] which attempts [illegible] the [illegible] levels. A. Alternatively, the pituitary may be [illegible] despite signals from the hypothalamus, and the target gland is also [illegible] B. The opposite may occur with gland [illegible] activity. The target gland [illegible] overactive despite [illegible] the pituitary. C. When this [illegible] overproducing, despite lack of signal from the hypothalamus, the target gland [illegible] to do likewise. D.



2 The thyroid gland in the [illegible] may [illegible] thyroid hormone [illegible] A. In a geriatric [illegible] enlargement [illegible] B. The method [illegible] radioactively [illegible] iodine [illegible] the uptake [illegible] into the [illegible] necessary for [illegible] thyroid hormone. Radioactive iodine is [illegible] small dose [illegible] the gland [illegible] by the dashes in the thyroid scans. C, D.

3 Body size may be affected by gland [illegible] disorders. In blood [illegible] A. Anterior pituitary overactivity [illegible] gigantism (as with the 7.34m (7.67ft) Dutch [illegible] overproduction of growth [illegible] is [illegible] of [illegible] as well [illegible] their companions.

Normally the thyroid gland is directly [illegible] by a hormone from the anterior pituitary gland at the base of the brain [1]. If the thyroid gland fails to produce sufficient thyroid hormone, this is registered by the hypothalamus (part of the brain) and the pituitary which produces thyroid stimulating hormone in greater amounts in an attempt to restore the normal balance.

Increased pituitary hormone production may, however, be a disorder in its own right rather than a response to failure of a target gland. If a tumour is present, for instance, hormone production may be excessive. Often only one of the several hormones that the pituitary produces is increased. The effects of [illegible] and abnormality vary according to which hormone is in excess and other body tissues and glands that respond [illegible] such hormone. If the pituitary is that which affects the adrenal cortex, part of the adrenal gland that lies just above the kidney. This hormone is adrenocorticotrophic hormone (ACTH). Excess ACTH leads to excess steroid hormone production by the adrenal glands. Disturbance of many systems results with retention of excess salt and water in the body and poor handling of sugar in the body. This disorder, called Cushing's syndrome, may result if the adrenal glands are overactive due to excess ACTH, but exactly the same pattern of symptoms arises if the adrenals are overactive in their own right.

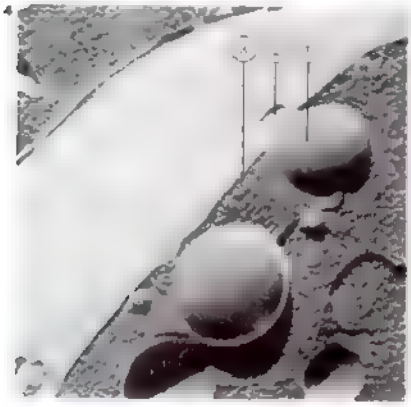
Other pituitary hormones

Disturbance of one hormone can have marked effects on normal feedback mechanisms, as is seen with growth hormone from the anterior pituitary. Its lack during childhood leads to dwarfism, while overproduction may lead to gigantism [3]. Too much growth hormone in later life leads to acromegaly. This disease is usually first noticed because the sufferer's hats, gloves and shoes need replacing with larger sizes. The nose, [illegible], hands and feet broaden and enlarge. Treatment is by X-ray irradiation.

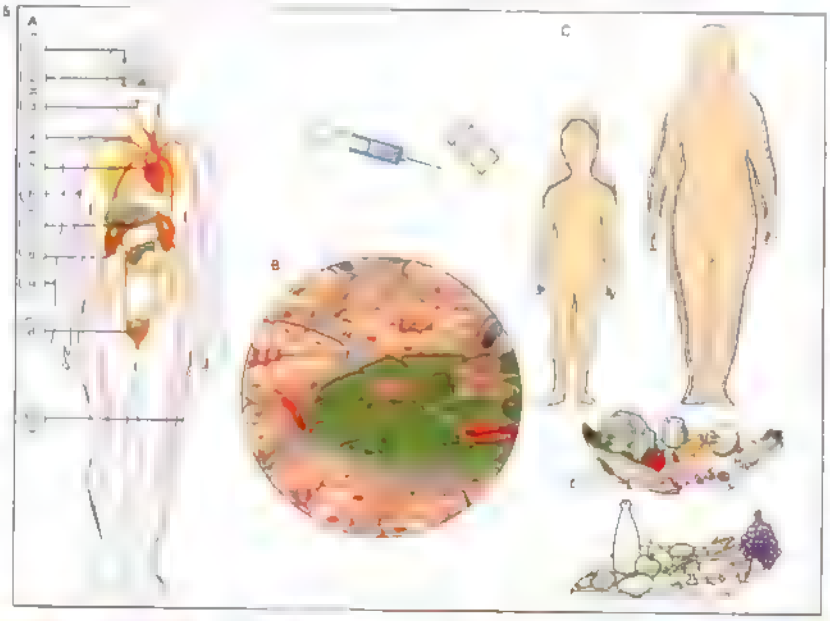
The posterior part of the pituitary normally produces antidiuretic hormone which maintains the correct amount of water in the body. If it is deficient, excess water is lost in the urine: this is diabetes insipidus.

KEY

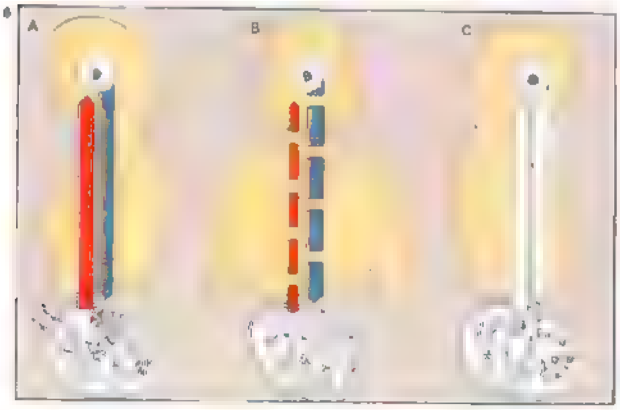
Daniel Lambert (1770-1809) weighed 335kg (737lb). He had a girth of 234cm (92in). The problem [illegible] justified by sufferers as being due to a glandular disorder. This is rarely true: over eating is the most [illegible] disorders of the body glands give rise to groups of symptoms that are fairly characteristic to [illegible]



4 Hormones are stored in the [illegible] and [illegible] specially prepared [illegible] throughout pan[illegible] When blood g[illegible] hormones are high after [illegible] through the cell wa[illegible] In diabetes insipidus [illegible] produced so low [illegible]

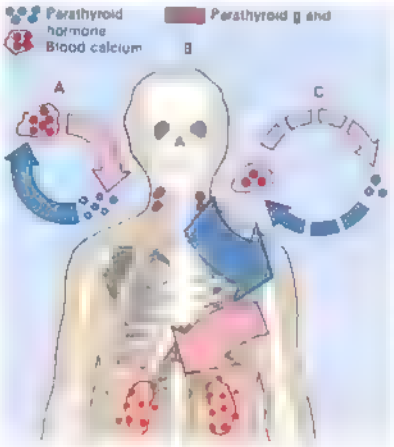


5 Diabetes mellitus A may [illegible] something [illegible] [illegible] 15 high blood sugar levels [6] fatty liver [7], kidney and [illegible] [8] [illegible] skin and delayed wound healing [9], and [illegible] weight [1[illegible]] [illegible] produced in the beta cells of the pancreatic islets of Langerhans (green, [B]). [illegible] ment [C] for juvenile [illegible] includes a varied low carbohydrate diet, usually in conjunction with regular [illegible] Adults [illegible] Low carbohydrate diet and weight loss [illegible] drugs may [illegible] control the disease [illegible] may be administered [illegible]



6 Before menopause A there is [illegible] feedback mechanism [illegible] which produce the hormones [illegible] and which also [illegible] at ng [illegible] change of [illegible] B which usually occurs at 45 to 50, the ovaries [illegible] this normal oestrogen [illegible] each month [illegible] result the pituitary produces an excess of hormones and the [illegible] and hot [illegible] are thought to be responsible for the common symptoms of sudden heat sense [illegible] headaches and weight gain. Oestrogens can be given as treatment. After menopause [C] hormone production ceases and the symptoms can disappear.

7 Overactivity of the four parathyroid glands in the neck leads to raised levels of calcium in the blood. A1 Calcium is also lost in the urine where kidney stones and kidney damage result. The bones are weakened as calcium and phosphate are removed from the bone structure. B1 in the body's attempt to maintain the high blood levels of calcium. Low levels [C] result if glands are damaged.



8 Apart from the endocrine [illegible] glands, [illegible] are other glands that do not secrete their products directly into the bloodstream but via ducts to the appropriate site. These exocrine glands are affected by their own disorders. A sufferer from mumps has a viral inflammation of the parotid gland. This is one of the salivary glands which produce saliva to moisten the food and begin food digestion. Mumps may affect other glands such as the ovaries and testes.

Diseases of the urogenital system

While the urinary and genital systems serve very different functions, they are often considered together for they share anatomical parts [Key]. The openings of these systems, the exits for urine, the penis in men and the vagina in women, are potential points where infections from outside may enter [1]. Symptoms may include difficult urination as in cystitis (inflamation of the bladder), blood in the urine [2], back pain and disturbances such as fever, sweating and general sickness

How infections start

Urinary infections are much more common in females, perhaps because the shorter urethra (the tube from the bladder to the outside) provides an easier path for invading bacteria than in the male Of the 300ml (0 6 pint) or so of urine in the full bladder all but 1-2ml are normally passed If bladder emptying is only partial, as may occur in the male with an enlarged prostate [4], any bacteria that reach the bladder are more likely to cause an infection as they are not adequately washed away. Congenital anatomical abnormalities present in the urinary tract, or the development of stones in the kidney, favour bacterial infections

Stones are formed from an accumulation of mineral substances filtered by the kidneys during the formation of urine Stones may cause intense pain (ureteral colic) if they block or move down the ureter If the kidney tubules, which are part of the kidney's two million or so filtering units or nephrons, are defective – either congenitally or as a consequence of pyelonephritis (severe kidney infection) – crystals of calcium salts may be deposited Stones may form also in patients with gout

The urogenital system can pick up infections during sexual intercourse These venereal disorders are of various types [5, 6, 7] They can produce both local symptoms at the time of infection and also permanent problems such as sterility. Other effects of syphilis may become apparent only after a period of years when widespread circulatory and nervous system diseases occur However, the advent of antibiotics allows these diseases to be cured when they are promptly treated by specialists Less serious but much more widespread of vaginal infections causing irritation and discharge These may be caused by fungi, especially *Candida albicans* (thrush) or microscopic amoeba-like organisms of the genus *Trichomonas*

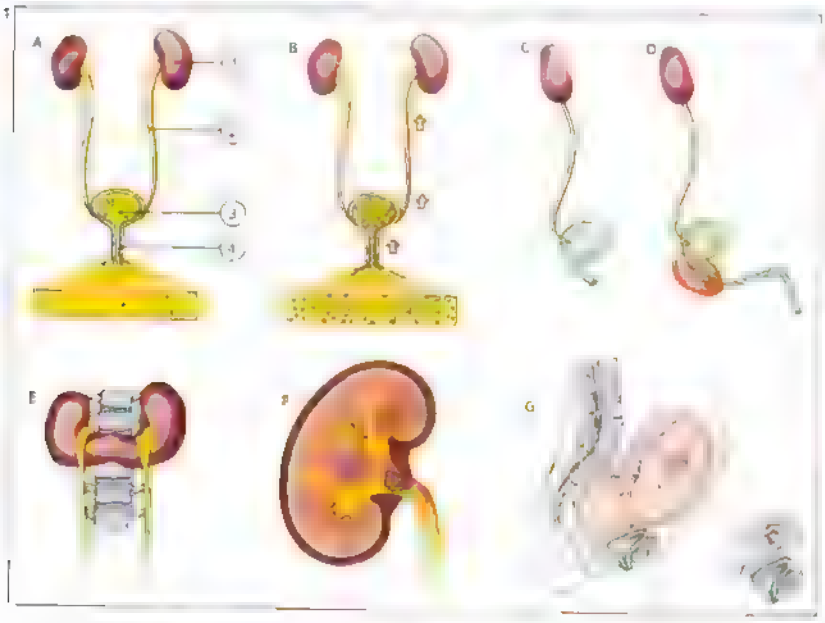
Sterility may be caused by venereal diseases and by other disorders of the urogenital system These disorders include the retention of one or both testicles inside the male body (the testicles normally descend to the exterior before birth) and malformation, damage or disease of the testes or ovaries that impairs the production of sperm and eggs

Kidney diseases

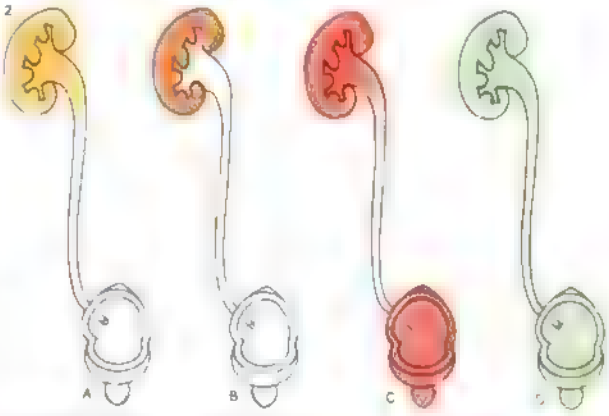
Kidneys may fail to function adequately either suddenly or slowly, and for many reasons. Less than 25 per cent of kidney tissue is necessary for normal removal of waste matter from the blood, and life is possible with even less than this If kidneys fail completely, it is now possible to use a machine [3] to do their job of removing waste material Alternatively, a diseased kidney can be replaced by a transplanted kidney. It is important that the transplanted kidney is

CONNECTIONS

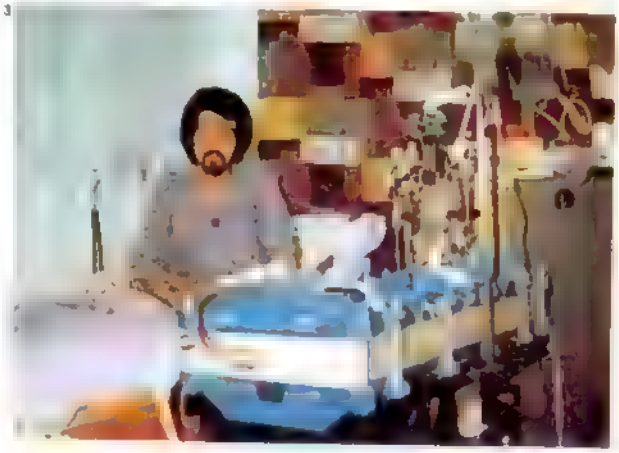
See also



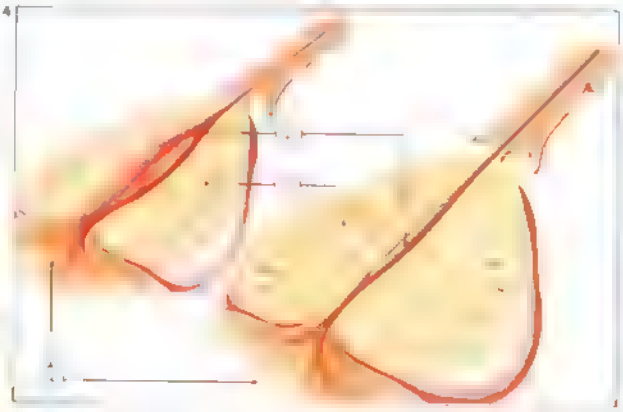
1 The urinary tract [A] consists of the kidneys [1], ureters [2], bladder [3] and urethra [4] Infection of the tract [B] first affects the urethra and bladder giving symptoms of cystitis with painful and frequent urination. If the infection spreads to the kidneys it causes inflammation and back pain A number of factors may lead to infection. The short urethra in the female [C] provides easy access for bacteria, while enlargement of the prostate in the male [D] may prevent the bladder emptying completely other inherited abnormalities of the kidney such as "horseshoe" kidney [E] and disorders such as kidney stones [F] predispose to urinary infection. More frequent urination during pregnancy [G] especially in the first two to three months and in the last month, is partly due to distortion of and pressure on the bladder, it encourages infection



2 Indications of diseases or infections of the urinary tract can often be found by routine examination of urine Abnormal contents such as blood pus and proteins can give some information on the type of disease and the part of the urinary system that is affected Kidney diseases such as glomerulonephritis [A] are indicated by protein in the urine. In serious stages [B] the protein coagulates into tiny cylindrical urinary casts. Damage to the kidney and bladder [C] is indicated by blood in the urine, and infection of the whole system [D] is indicated by the presence of bacteria in the urine



3 A kidney dialysis machine is used if nearly all kidney units or nephrons fail, bringing the danger of death from an accumulation of waste products in the patient's blood The patient can be connected in hospital, or later at home, to a filtering system that removes the impurities Two or three overnight sessions a week are usually required to pass the blood with its impurities through a solution that has the correct concentration of salts



4 The prostate gland [1] is a partly glandular firm structure that circles the male urethra [2] at the base of the bladder [3] It produces some of the seminal fluid in which sperm are ejaculated With increasing years, the gland enlarges, normally a harmless process However, it may constrict the urethra and prevent adequate urination Urine becomes increasingly slow and difficult to pass with dribble at the end of the stream Infection can occur the kidneys may be damaged and complete and painful stoppage of the flow may result Treatment is usually required when this happens and a so-called urination is too frequent This involves surgical removal of at least part of the gland

carefully matched (like the matching of blood before transfusion), but drug treatment is necessary to help prevent the body's rejection of the graft

Sudden kidney failure may occur if the blood pressure and flow through the kidneys drop too low – as after haemorrhage or severe burns. It is also a feature of acute glomerular disease where extensive swelling and inflammation is found in the glomeruli, which are parts of the filtering units. This can follow bacterial infection of, say, the throat by streptococci

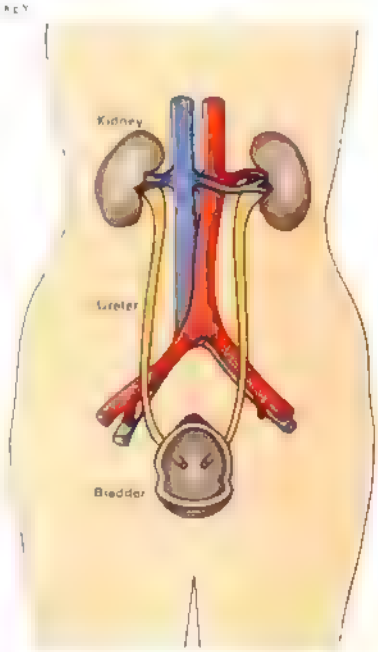
Many disorders can result from chronic renal (kidney) disease in which the ability of the kidney to remove the waste materials from the blood gradually deteriorates. The breakdown products of proteins are dealt with particularly badly by the failing kidney, and limitation of protein foods may be an essential part of the medical treatment.

Types of tumours

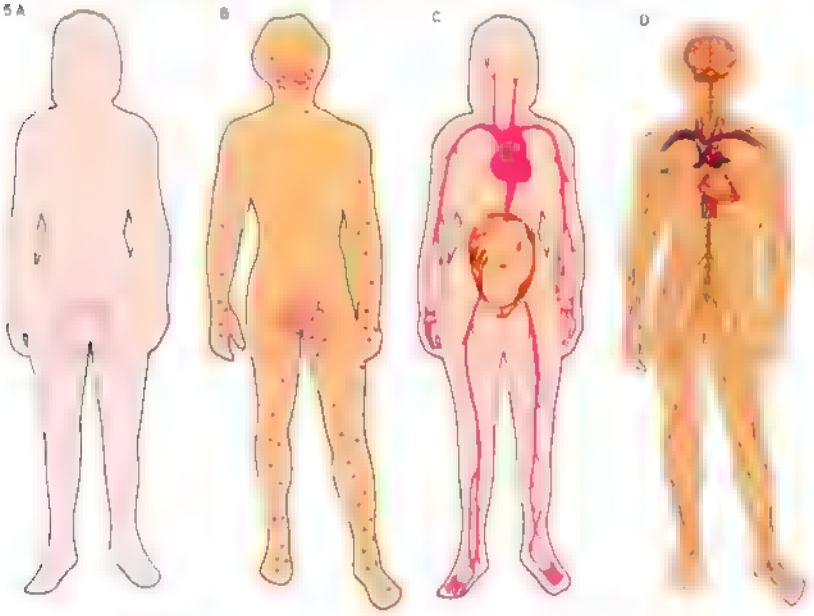
Like other parts of the body the urogenital system is at risk from tumour formation In children, several rare types of kidney cancer are usually malignant, while in later life bladder tumours may be benign or malignant. They may be single or multiple polyps (wart-like growths) and can cause blood to appear in the urine. Often polyps can be removed or burned through a cystoscope – a tube inserted through the urethra by which the bladder can be inspected and treated

The genitalia and reproductive organs may also show tumour growth Tumours of the testes are rare, but the female reproductive organs are common sites for both benign tumours, often called fibroids, and malignant ones known as cancers. It has been found that cancer of the body of the uterus (womb) is more likely in women who have had no children Similarly breast cancer is a little more common in those women who have no children, or in whom the first child was born when the mother was older than 30

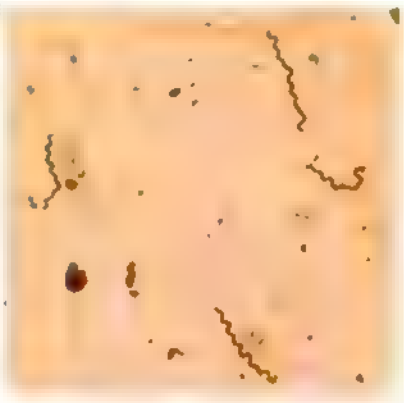
In contrast, cancer of the cervix, or neck of the womb, seems to occur in women with several children, and its incidence is influenced by factors in the male Circumcision of the penis, or careful hygiene in the uncircumcised, appears to reduce the risk

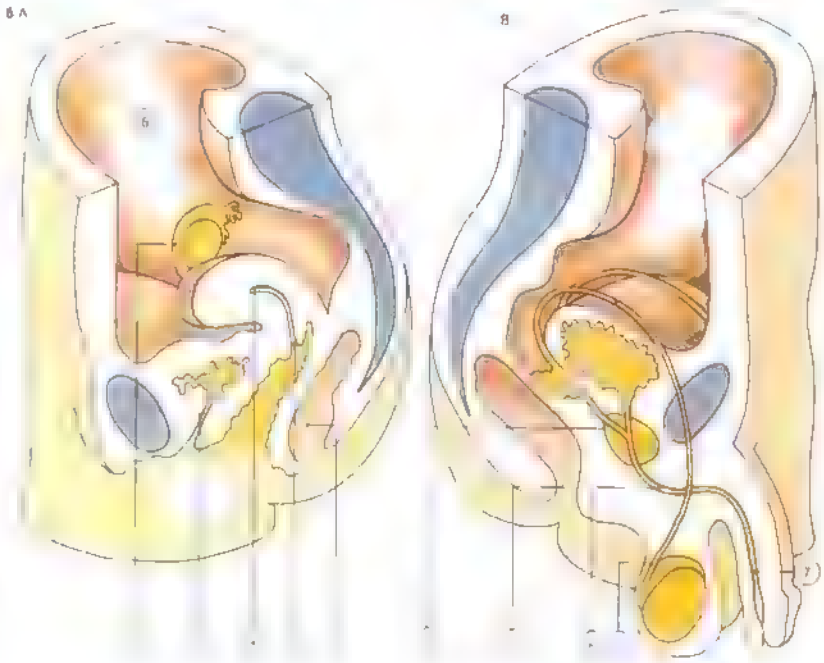


The urinary system consists of two kidneys and ureters draining into a bladder and urethra (urine duct) Its function is mainly to produce and excrete urine, in this way maintaining the water and chemical balance of the body Disorders of or damage to the kidneys may prevent normal urine production, while an enlarged prostate gland in the male may prevent the free passage of urine down the urethra Stones, often of calcium salts, may form in the kidneys, while tumours may grow in the urinary tract As in other tissues, infection can occur The system is rather more vulnerable than some as the exit for urine may be an entry for bacteria

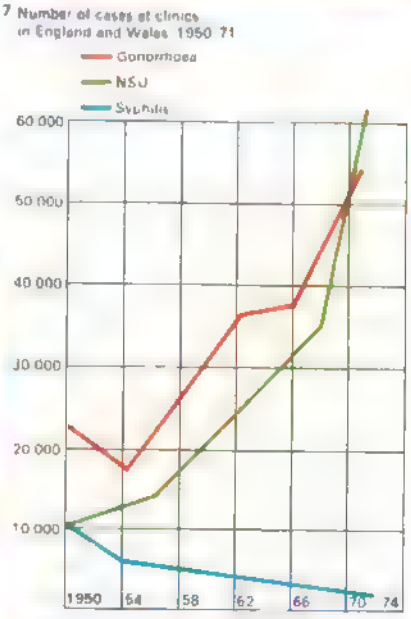


5 Syphilis is an infectious disease transmitted during sexual intercourse Generally a sore or chancre appears painlessly at the site of the infection (A) usually on the genitalia, within ten weeks The next phase occurs with a spread of bacteria by means of the blood causing mild illness with skin and mucous membrane rash (B). There may be no symptoms after this but a fetus in the uterus may be infected and born with congenital syphilis (C) The disease may be dormant for up to 40 years but late effects can extensively damage heart and blood vessels, nerves and brain (D) The bacterium is a spirochaete (E)





6 Gonorrhoea, a bacterial disease usually transmitted by sexual intercourse causes inflammation of the urinary and reproductive organs In females (A) it causes mild inflammation of the urethra (1), cervix (2) and sometimes the rectum (3) The bacteria may spread to the uterus (4) and the Fallopian tubes and ovaries (5), causing infertility Symptoms elsewhere may follow, spread by the bloodstream (6) In the male (B), similar effects occur but are usually more acute with pus discharged from the urethra (7) which may narrow higher up (8) The rectum (9), prostate (10) and epididymis (11) may be infected through the blood [illegible] for this essential



7 Cases of venereal disease increased sharply in 1950-74 and are still rising This cannot be accounted for simply by more open discussion of the condition and greater willingness to come forward and be treated The rise may be related to increased promiscuity and the fact that oral contraceptives may decrease the naturally protective acidity of the vagina Incidents of gonorrhoea and non-specific urethritis (NSU) have increased more than syphilis

Alcoholism and drug abuse

In every known society some people have formed the habit of taking substances that cause alterations of consciousness. There are two main groups of such substances, alcoholic drinks such as beers, wines and spirits, and drugs such as cannabis, mescaline, cocaine, heroin and LSD. Some of these are relatively mild and harmless, except when taken in repeatedly large doses, others, even in small doses, set up a dependence that can totally dominate the individual.

Susceptibility to alcohol

Alcoholics are drinkers who depend on alcohol to such an extent that eventually they show noticeable mental or physical disturbance. Most alcoholics are persistent heavy drinkers [1] but various patterns of alcohol abuse have been identified. Some drink steadily over many years and suffer physical damage in late middle age as a consequence. Others can function quite well without alcohol for quite lengthy periods but when exposed to alcohol cannot control the amount they drink. Bouts of heavy drinking lasting several weeks are not uncommon.

Most alcoholics are middle aged and male although in recent years there are signs that more young people and more women are abusing alcohol. People in certain occupations, such as waiters, salesmen, seamen, company directors, barmen and people working in the liquor industry, are particularly liable to alcoholism.

Alcoholism usually develops slowly. The alcoholic often begins by relying on alcohol to ease anxiety or depression, which it does briefly. However, tolerance to alcohol develops rapidly and increasing amounts are drunk to obtain relief. Gradually, the alcoholic develops physiological dependence on alcohol so that when deprived of it he exhibits withdrawal symptoms [3]. These include nausea and vomiting, tremors ("the shakes"), memory lapses, epileptic fits and delirium tremens [4]. Pathological drinking is invariably accompanied by difficulties at work due to absenteeism or drunkenness, family and marital disharmony, financial difficulties and mental and physical ill health.

Physical complications of alcohol abuse include gastritis, peptic ulceration, cirrhosis of the liver, inflammation of the pancreas and damage to the heart muscles. Brain damage with severe memory impairment is not uncommon. Psychiatric complications include severe depression with marked guilt feelings which sometimes ends in suicide.

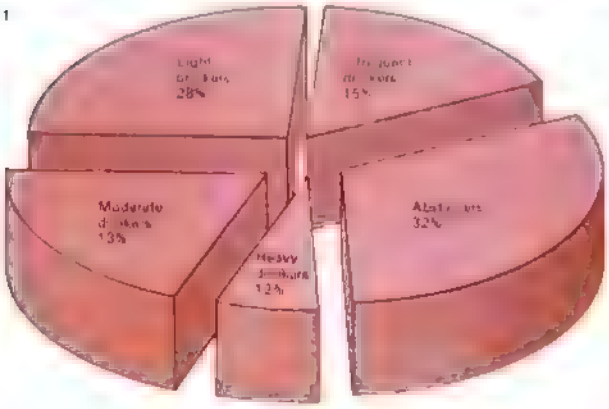
Treatment of alcoholism is aimed at developing an awareness on the part of the alcoholic that he has a problem with drink. Some experts believe there is a form of metabolic fault in alcoholics and complete abstention seems to be necessary in most cases. Psychotherapy, aversion treatment and drugs are used to help achieve this.

The dangers of drug dependence

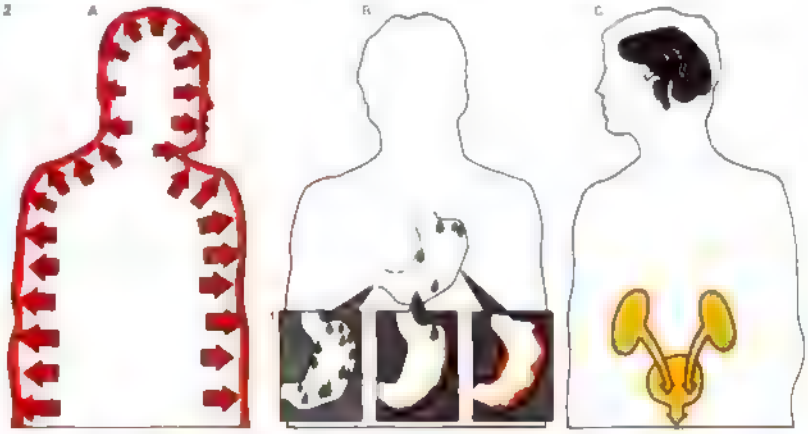
Drug dependence means the repeated non-medicinal use of a drug causing harm to the user or to others. All drugs can be dangerous and should be treated with extreme caution, irrespective of some current social attitudes that tend to glamorize the use of certain drugs. Drugs have been produced from a bewildering variety of plants but a large group, including morphine and heroin, are derived from the poppy and are called

CONNECTIONS

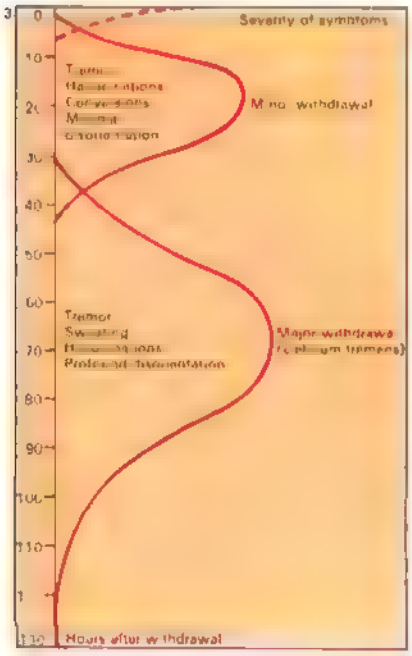
See also



1 Drinking categories in a 1969 American survey assumed that abstainers drank less than once yearly, infrequent drinkers less than once monthly, light drinkers at least once monthly but not more than a drink (pint of beer or single measure of spirits) per session, moderate drinkers at least once monthly, three-four drinks per session, and heavy drinkers almost daily, with up to five drinks per session.



2 Short-term effects of alcohol on the body include dilation of skin blood vessels [A] and a drop in temperature. Gastric secretion [B] is increased at a low concentration (20mg alcohol/100ml blood) [1] but is inhibited at higher concentrations [2] and the stomach lining is irritated [3]. Inhibition of the pituitary gland in the brain causes increased production of urine [C]. Vital centres in the brain are depressed [D]. Judgment, self criticism, perception and motor skills are impaired and sleepiness, coma and even death may occur [E]. Drinkers may change from being carefree to being irritable, disorderly and, finally, comatose [F].



3 Withdrawal symptoms can follow cessation of drinking after a heavy or chronic intake of alcohol. They can be relatively slight, occurring within 8–48 hours of withdrawal, or quite severe, extending over a period of three to eight days.

opiates [Key] New addictive drugs have recently been developed in laboratories.

Addicts who cannot finance their habit (the rapidly developing tolerance to opiates means that ever-increasing doses are required) often turn to crime to finance their addiction. Deprived of his "fix" (jargon for injectable dose), the addict experiences severe restlessness, vomiting, diarrhoea and insomnia, eight to twelve hours after his last dose. Such symptoms last between three and seven days and can cause death. Complications of drug addiction include overdose, liver damage, sepsis and pneumonia.

Dependence on the so-called soft drugs (the barbiturates and the minor tranquillizers) is a serious problem, given the popularity of these drugs in the treatment of insomnia [5] and anxiety. Such drugs differ in their ability to produce tolerance and physiological dependence, but their use, depending on the personality and social pressures involved, can in time lead to the abuse of other drugs, such as the opiates

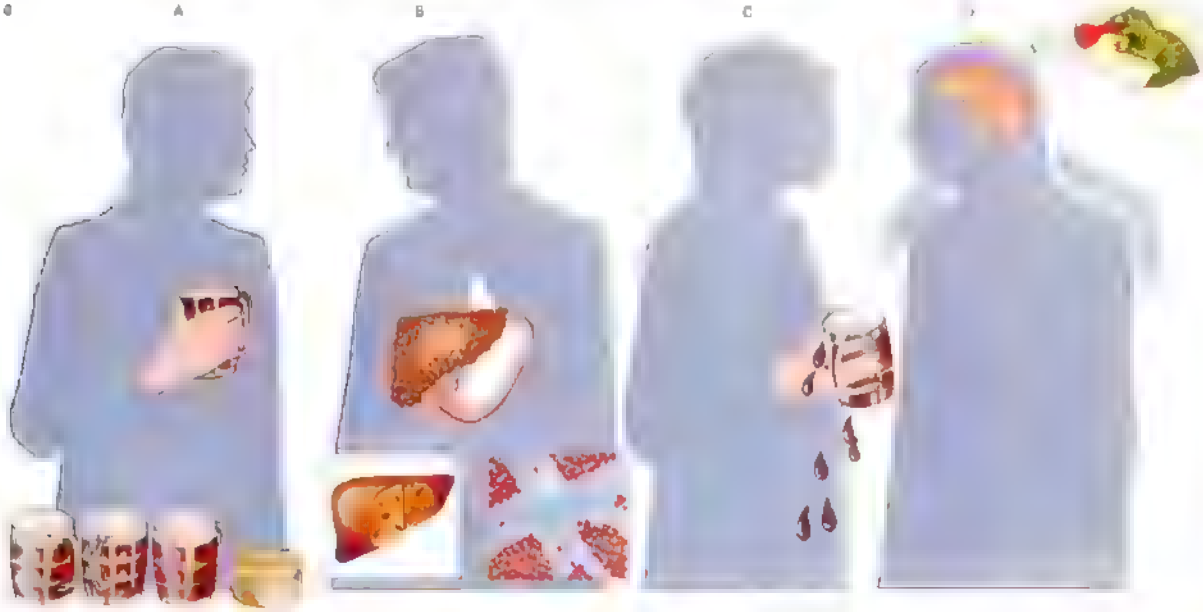
Illusion, hallucination, altered time sense, distorted judgment, confusion and disorientation are experienced in LSD "trips". There have also been reports of flashbacks – recurrences of the drug effects weeks or even months after the end of a trip No withdrawal symptoms occur after taking LSD nor on stopping amphetamine intake Amphetamines [6] produce a rapid onset of euphoria and in large doses can cause a severe paranoid psychosis.

The cannabis controversy

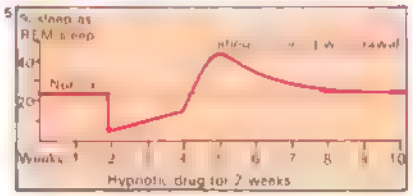
Whether *Cannabis sativa*, a euphoriant and relaxant derived from Indian hemp, is a dangerous drug of addiction or a harmless pleasure remains controversial. The dried leaves of the hemp are termed marihuana, the resin obtained from the flowering tops, hashish. In moderate doses, cannabis can produce a sense of excitement, heightened awareness and well-being followed by a phase of tranquillity and then fatigue

Drug dependence can be treated by maintaining the addict on a controlled dose of his drug, by replacing the drug by a similar but less potent synthetic or by controlling withdrawal under supervised conditions

Opium smoking was fostered in China during the 19th century by European traders intent on profit, particularly the British who crushed a Chinese attempt to stop traffic in the drug By the mid-century, so-called opium dens were sordid features of Chinese cities Although Europeans were less addicted, opium, laudanum and morphine were legal and common drugs in Europe as well as America in the same period

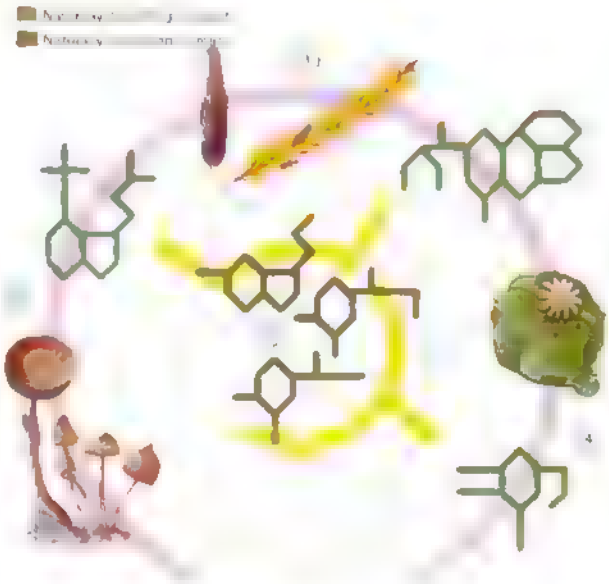


4 People who drink heavily often eat poorly [A] and, in the long term, malnutrition can result Inflammation of the stomach and disease of the liver (cirrhosis) produce further loss of appetite. In cirrhosis [B], the liver is damaged either by the direct toxic effect of alcohol or by its effect on nutrition Normal liver tissue is replaced by fibrous scar tissue Deficiency of vitamin B1 damages the long nerves to the limbs, resulting in peripheral neuritis [C] A sufferer loses touch sensitivity Delirium tremens [D], characterized by extreme agitation and visual hallucinations, is a serious withdrawal effect



5 Sleeping tablets reduce the amount of REM (rapid eye movement) sleep when dreaming occurs But REM sleep is higher than normal after withdrawal of the drug, with effects akin to delirium

6 The amphetamine group of stimulant drugs acts on the brain in the region of the reticular formation and hypothalamus (yellow) They can be taken as tablets or injected intravenously to decrease fatigue, increase alertness and lift mood. Tolerance develops rapidly, making progressively larger doses necessary. The euphoria, which follows injection particularly, is short-lived and gives way to depression Severe paranoid states can occur



7 The hallucinogenic drugs may produce their effects because they are similar structurally to substances [1] thought to be neurochemical transmitters. Three of the most common naturally occurring hallucinogenic substances are psilocybin, LSD and mescaline Psilocybin [2] is extracted from the psilocybe family of mushrooms LSD [3], a highly potent hallucinogen used extensively for therapy by some psychiatrists, was first isolated from a fungus, *Claviceps purpurea*, which causes ergot of rye Mescaline [4], which has played an important role in the religious rituals of certain Mexican Indian tribes, is obtained from the peyote cactus

Diseases of the Third World

Health or lack of disease depends among other things, on quantity and quality of diet effective control of housing, food standards and water and sewerage systems, and the availability of doctors, drugs and hospitals should a disease occur This is a measure of the problem facing most of the developing countries that make up approximately 70 per cent of the world s 4,000 million population All these factors are expensive and presuppose a solid industrial base – something that many Third World countries lack Hence their health problems are immense: the life expectancy of a newborn baby in such a country is still about 35 years – about the same as it was in Western societies in the fourteenth century

Insect-borne diseases

The diseases that still affect so many people fall into two main groups infectious disorders and nutritional ones The most important group of infections that occur in tropical countries are those spread by insects, which carry disease from person to person

First among these insect-borne diseases is malaria, still one of the world's greatest killers. It is caused by a tiny parasite that is carried from man to man by the *Anopheles* mosquito [1]. When a mosquito sucks blood from an infected person the parasites are sucked up into the insect's stomach along with the blood They breed and ten days later their offspring can be found in the mosquito's salivary glands From then on the insect will inject a dose of parasites into anyone it bites

The main symptoms of malaria are high fever, headache and violent shivering (rigors) While some species of the malaria parasite are relatively benign others often cause chronic ill health or death Treatment is with drugs such as chloroquine, but prevention is a better solution There are three types of prevention killing the mosquitoes themselves [1], stopping them from biting people by using sleeping nets and repellants, and the administration of protective drugs

Other mosquito-borne infections include various types of filariasis [3], common in the Pacific, the Far East and Africa and yellow fever, a serious virus disease which occurs in West Africa and South America For many years now yellow fever has hindered the development and cultivation of much of the South American hinterland Prevention is the same as for malaria but immunization is an added precaution

Several other insects are disease carriers Leishmaniasis [4] and phlebotomus fever are all spread by sandflies and cause much suffering and economic disruption, especially in the Middle East Rat fleas still continue to spread bubonic plague [5] in some Third World countries while the tsetse fly is responsible for trypanosomiasis better known as African sleeping sickness [6] or South America's Chagas' disease

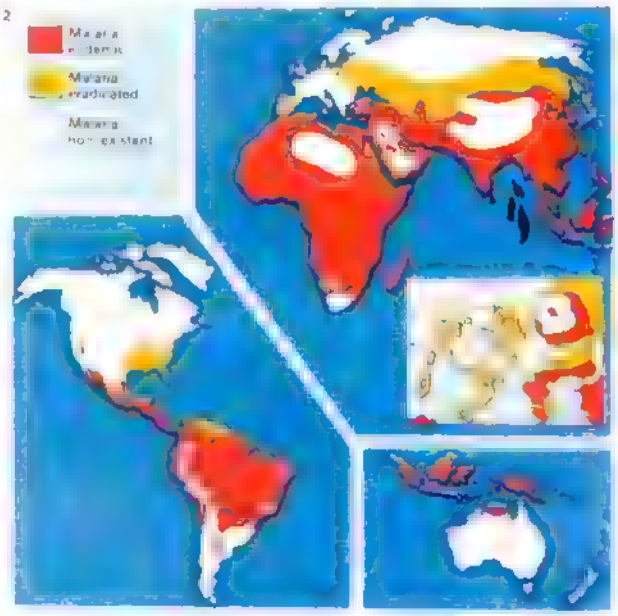
Lice are the carriers of ordinary typhus, while tick typhus is spread by dog ticks, and scrub typhus (tsutsugamuchi) is carried by rat mites, rat fleas carry murine typhus

Diseases of insanitation

Almost as important as the insect-borne diseases are those of insanitation These depend for their spread on the contamination of drinking water or food by human faeces The linked problems of disposing of excreta and

CONNECTIONS

1 The malaria parasite is carried by the *Anopheles* mosquito which thrives in areas where stagnant water is common The female lays her eggs in water [1] from them larvae [2] develop These go through a pupal stage [3] before developing into adult mosquitoes [4] Insecticides and [illegible] are preventives

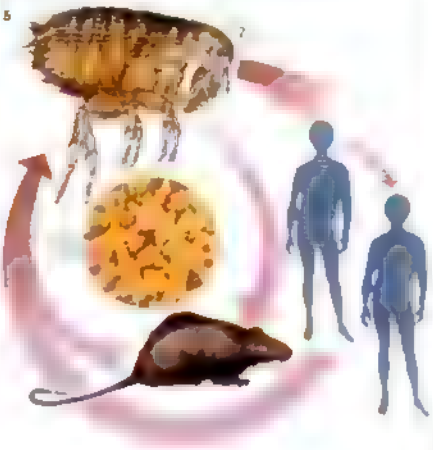


2 Progress in eradicating [illegible]

3 Elephantiasis is one of the most dreadful types of filariasis which is usually caused by the minute filarial worm *Wuchereria bancrofti* These worms are injected into the human bloodstream by the bite of an infected mosquito

4 Leishmaniasis is a widespread disease of tropical countries caused by a tiny parasite that is spread by infested sandflies When the parasite enters the bloodstream [1] it changes form [2] and is ingested by a white blood cell [3] The parasite multiplies in the cell and its descendants may attack the skin [4] or lymph nodes [5] If the victim is bitten by another sand fly, parasites [6] will be sucked up by the insect with the blood and mature into forms that can infect other humans [7]

5 Bubonic plague ("Black Death") may still occur especially in the Indian sub continent It is caused by the germ *Pasteurella pestis* [1] which can live in man in rats and in rat fleas [2] If a rat has the infection the fleas that feed on it will ingest the germ A man bitten by these fleas will probably develop the plague – and may spread the infecting organism to other humans via his breath



of providing pure drinking water are both difficult and costly to solve

Infections that are spread by this "faecal-oral route" include many kinds of non specific diarrhoea as well as typhoid paratyphoid, cholera, various types of food poisoning, bacillary and amoebic dysentery and probably polio as well. Of this group the disease causing most concern in the 1970s was cholera [7], which was spread alarmingly from its base in the Far East to much of the world in those years. It also penetrated south of the Sahara desert for the first time in medical history

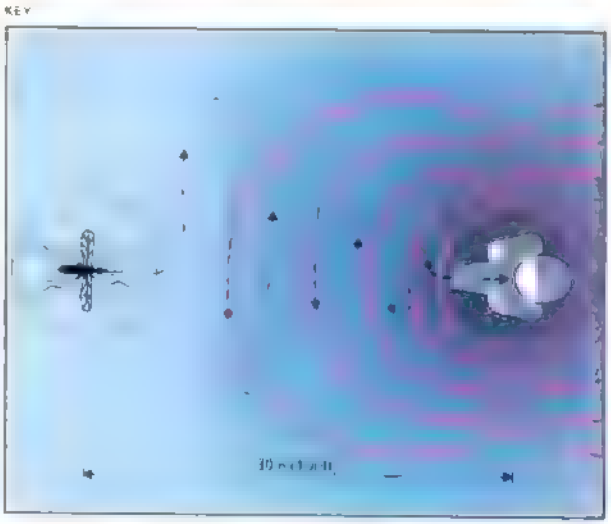
Other sanitation-related infections to which the Third World populations succumb include hookworm and other worm infestations, which produce anaemia and chronic invalidism, and the tropical disease bilharzia or schistosomiasis [9]. The economic effects of bilharzia, a crippling disease that drastically reduces the ability to work, are severe

Less important numerically are leprosy and yaws (both spread by prolonged personal contact only, and both now treatable) and rabies which is caught from animals, especially dogs, and is almost untreatable. The last of the important infectious diseases is smallpox, although this seems, at last, to be well under control and is confined to only six countries in Asia and northeast Africa

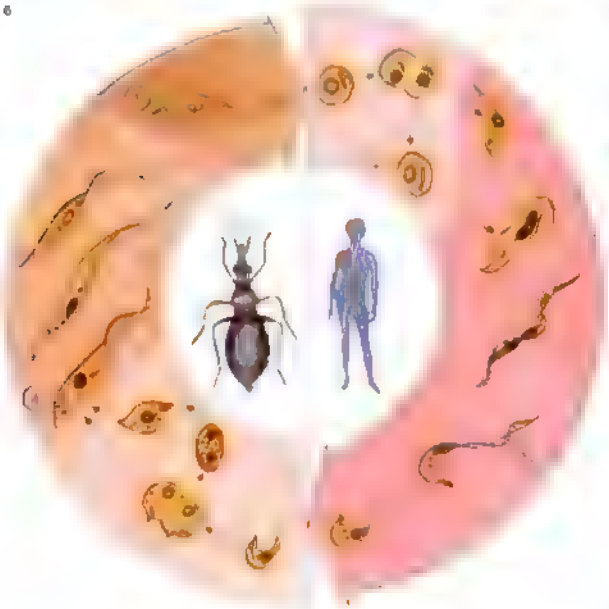
Nutritional disorders

The well-known vitamin deficiency diseases such as pellagra, beri-beri and scurvy are not now of enormous importance in the Third World, partly because it is relatively easy to provide the very small quantities of vitamins needed to maintain health

Far more serious are outright starvation and the related disease of kwashiorkor or protein starvation. Half the world still goes hungry and in areas such as North Africa, where crops have repeatedly been hit by drought, or Bangladesh where floods have destroyed food supplies, millions of people find it desperately hard to obtain any kind of food. As the world's population grows inexorably and as the amount of protein available per mouth gets less and less, it seems that the problem of starvation is likely to get worse rather than better

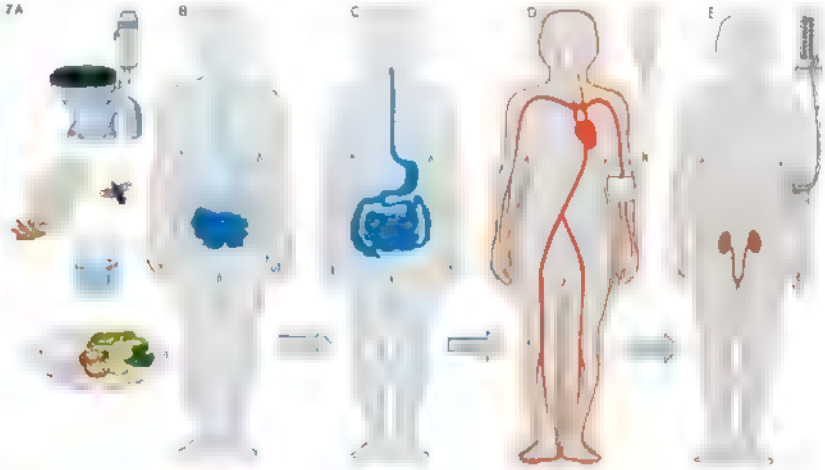


Mosquitoes find their victims in two stages. Carbon dioxide, which humans exhale, first sets them flying. If they then sense a current of warm, wet air, such as every person gives off, they home in on it, some people are more prone to mosquito bites than others because they are warmer and wetter. Repellants interfere with the accuracy of the mosquitoes' sensors



6 Trypanosome infections cause African sleeping sickness and South America's Chagas disease, which often produces fatal heart damage. The parasite [1] enters the blood to lodging a bite by an infected insect such as an assassin bug or tsetse fly, it reproduces [2] and develops into forms that can be ingested by other bugs [3]. In these bugs intermediate forms [4] develop into parasites to infect more people

7 Cholera is one of the many diseases that spread because of poor hygiene. Human faeces containing cholera germs contaminate food or water [A]. The victim ingests them and they incubate in his small intestine [B]. They then produce severe often fatal diarrhoea [C]. The patient may survive if his gross dehydration can be corrected with intravenous fluid [D]. If such therapy is undertaken promptly (and if no kidney damage has occurred) then full recovery should take place [E]



8 Insanitary conditions associated with food, water supply or sewage disposal are a sure way to spread disease. This Afghanistan meat market perched over an open sewer, could be a source of infection

9 Bilharzia (schistosomiasis), named after Theodor Bilharz who identified it in 1851, today affects more than 200 million people in tropical countries. It is caused by a minute parasite that spends part of its life cycle in freshwater snails [1] and part in water when it can penetrate the skin of anyone washing or bathing [2]. The invading parasite lays up to 40,000 eggs after developing into an adult. These interfere with blood flow and damage liver, lungs and kidneys.

10 Leprosy has been one of the most dreaded diseases for many centuries, yet it is not highly infectious and efforts to control it are now meeting with great success. Caused by a bacterium (Hansen's bacillus) it is thought to be passed on only by prolonged close contact. It mainly attacks the skin and the nerves and may produce either trivial alterations in the tissues or gross deformities

World health

The World Health Organization (WHO) defines "health" as "a state of complete physical, mental and social well-being, and not merely the absence of a disease"

The object of WHO is to try to achieve such a state of well-being for as many as possible of the world's peoples. It may seem at the moment that such a target is far off, when so many millions of people are suffering from disease and malnutrition. But in fact today's pattern of health and disease is in many ways far better than that of past centuries. In Western countries at least, men and women are living longer and healthier lives than ever before – even though "warning signs" in the shape of diseases of "civilization" (such as lung cancer, psychological stress and coronary thrombosis) are now appearing with much greater frequency.

Health problems in the Third World

Even in the poorer parts of the world there have been great improvements in health care during this century Immunization against infectious diseases has helped to control once-dreadful scourges such as smallpox [7], and energetic measures against the mosquitoes, flies and ticks that carry disease are slowly helping to defeat killer infections such as malaria, yellow fever and typhus

The health problems of the poorer two-thirds of the world are still far from being solved, however – a fact easily demonstrated by the mortality figures of various nations [4]. In the less well-off countries there are high mortality rates among the young (in Nigeria over 180 babies out of every 1,000 die in their first year) and survival into later life is still the exception rather than the rule [1]

Several factors account for this situation One is that the tropical climate of many developing countries encourages the development and expansion of the organisms that cause life-threatening infectious diseases. Another is that providing and maintaining good health care for a rapidly expanding population poses extreme economic problems. In much of the world infant mortality is high simply because there are no proper facilities for looking after mothers and their babies. Those babies who survive infancy may well die in childhood because of malnutrition. And those who achieve adulthood may well die young because of lack of adequate medical care during illness or maternity. People who are weakened by malnutrition, and who have no doctor within reach, are easy targets for any kind of disease, especially the virulent tropical infections. These diseases need not be confined to the tropics. Today's air travel means that such infections can be carried to all parts of the globe in a short time Monitoring the spread of such diseases is one of the major challenges to world health

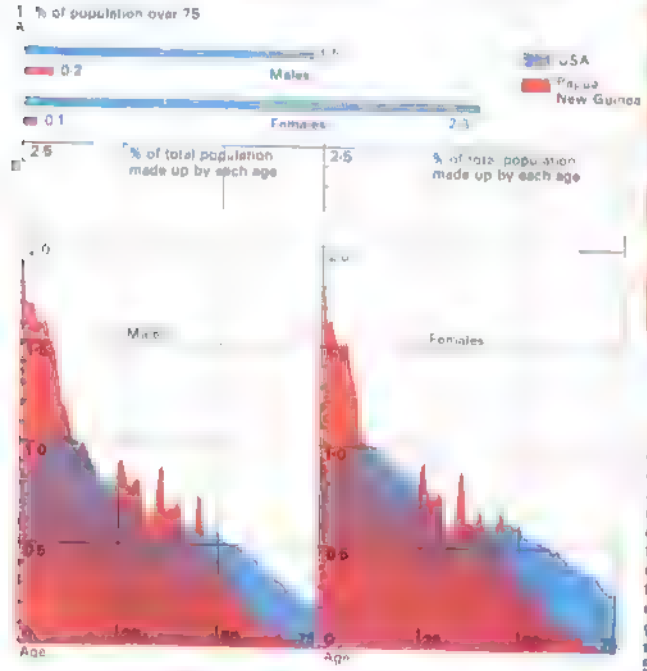
Organizations to the rescue

It was to deal with all these difficult and complex problems that the United Nations set up the World Health Organization in 1948 [Key] By the mid-1970s WHO was active in more than 130 different countries and cost $1,000 million a year to run

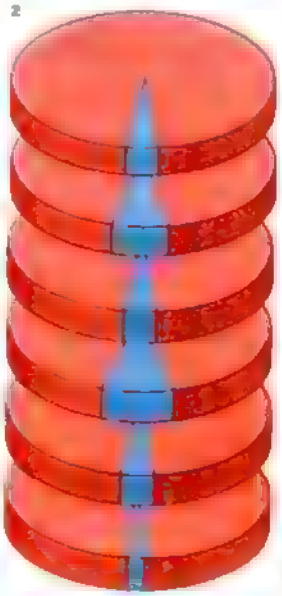
The functions of WHO are to act as a clearing house for medical information, to carry out research (especially in the field of epidemiology, whereby the origin and spread of disease is monitored by the Epidemiolog-

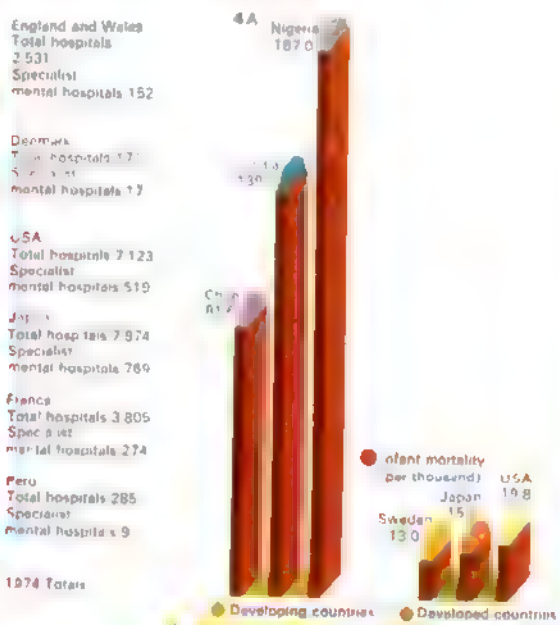
CONNECTIONS

See also

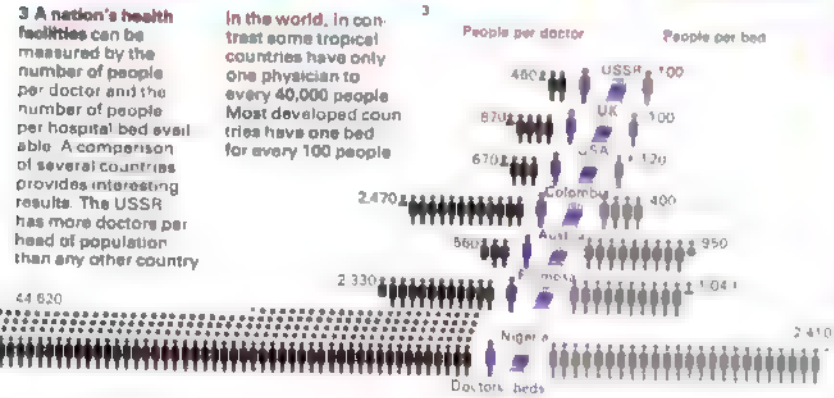


1 Developed and developing countries vary greatly in the availability of health care and this is reflected in the average age of their populations Comparison of the percentage of the populations of the USA and Papua New Guinea over 75 [A] or of their general age distribution curves [B] shows marked differences. In developing countries with high birth rates and low life expectancy the average age of the population is often below 20 and the number of persons living into the 40s and 50s is relatively small In developed countries most people survive early childhood and the average age is much greater





2 While physical disease gradually declines many types of mental illnesses are increasing, especially in industrialized countries. Despite the importance of mental health the total number of specialist mental hospitals giving constant, skilled attention is 5–10% of the whole



3 A nation's health facilities can be measured by the number of people per doctor and the number of people per hospital bed available A comparison of several countries provides interesting results. The USSR has more doctors per head of population than any other country in the world. In contrast some tropical countries have only one physician to every 40,000 people Most developed countries have one bed for every 100 people

4 National wealth is the key to good health, as found in developed countries. It can secure balanced diets, good living conditions and expert medical attention High infant mortality and reduced life expectancy are indicators of poor health and are common in developing countries where such benefits are not easily available. In Sweden, Japan and the USA, fewer than 20 babies in every thousand die during the first year of life In Nigeria the figure is nearly 200 [A] Similarly, life expectancy [B] in the same three developed countries is over 70 years, but in Nigeria a newborn baby can only expect (on average) to live to 37 Much of Africa and parts of Asia have similar low life expectancy levels

ical Intelligence Service), to administer international sanitary regulations, especially in connection with travel and quarantine, to finance international research programmes including studies of cancer and tuberculosis and to help individual countries train medical staff and to fight disease [9]

In this last field WHO does much to provide vaccines, antibiotics and other medicines It also helps to disseminate health education information, which is of great importance in countries where disease and overpopulation are often due to ignorance

The question of overpopulation now looms large in current thinking on world health It is generally recognized that there are too many people in the world for its nutritional resources and health care facilities to cope with. More and more people are going to starve to death (or die from lack of medical treatment) unless the earth's rapidly burgeoning population can be limited Much current world health work is devoted to fertility control research and to teaching people about the urgent need for family planning

Part of the solution of present problems will lie in increased food production and in this field WHO works closely with one of the UN's other semi-autonomous bodies, the Food and Agriculture Organization (FAO), which has done much to improve world health by research into new crops and new fertilizers and by educating people in better techniques of farming. Unfortunately it seems almost certain that whatever scientific breakthroughs are achieved in this field there will never be enough food to feed even the world's present population

Hope for the future

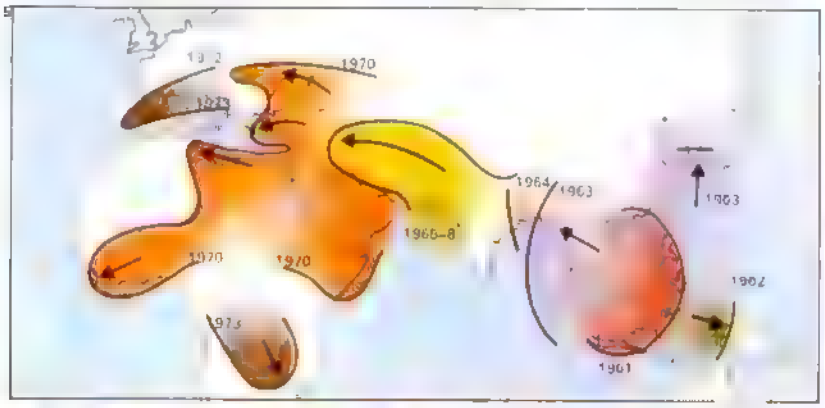
If there is to be any hope of a future for the Third World renewed efforts must be made to eradicate the dangerous communicable diseases of the developing world and to prevent the spread of the "civilization diseases of the West People must be educated in the ways of good health and more health personnel must be provided More food must be grown. If all this can be achieved then there is some hope that the advances of this century may be shared by a far wider proportion of the world's population than at present

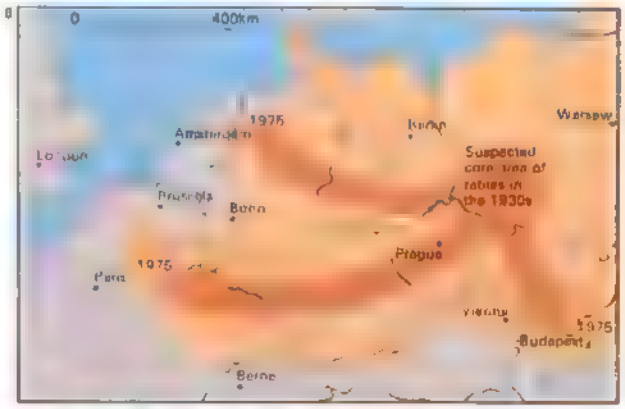
KEY

The World Health Organization (WHO), whose symbol is shown here, is a semi autonomous unit within the United Nations Organization It is based in Geneva Switzerland, and is active in more than 130 countries, mostly in the Third World

5 Cholera spread in the early 1960s from its "base" in the Far East with alarming rapidity Some cholera germs had changed their nature and this new variant was called the "El Tor" type Although cholera has not been seen in most Western countries since mid-Victorian times, by the mid 1970s, as the map shows, it was breaking out in southern Europe, aided by increased air travel

6 The spread of rabies (hydrophobia) across Europe is shown on this map The actual incidence of rabies is unknown WHO reports 600 to 700 deaths a year – but the figure is probably far higher The virus of this terrible disease is spread by dog bites (although even a lick from an infected animal may prove fatal) Foxes and other wild animals can also be carriers of the disease





8 The thin, wasted muscles and protruding bones of these beggars in Calcutta, India are almost a badge of the developing countries and a sign of the tremendous problems facing them These problems are exaggerated by the natural disasters (typhoons, flooding and droughts) that seem to plague them regularly

7 Smallpox, caused by a virus that is spread by contact with an infected human, was once a world scourge that killed or disfigured thousands of people (even in the most economically advanced countries) every year Now it has been driven back so nearly to total extinction that many Western countries have abandoned routine vaccination of infants against it. Travellers may still need immunization, however

9 Regular, small-scale clinics, like this one at Keneba Gambia, for fighting malaria contribute significantly to the control of disease Particularly in areas of poor transport and few ancillary services such as electricity, this type of centre tailor made for local conditions, is more effective than prestigious hospitals of the Western type

Community medicine

Anyone who has suffered from an attack of food poisoning after a meal in a restaurant, caught influenza from someone on a bus or fleas from a hotel bed or developed athlete's foot after visiting a public swimming pool will know only too well how the health of the individual depends on the health and cleanliness of the whole community

People cannot function properly when they are ill and in the same way a society cannot function properly unless its members are healthy. As a measure of self-protection, therefore, every community takes steps to promote the health of its members. This is increasingly important as the size of the community increases, both because the results of ignorance, negligence or incompetence can potentially affect many more people more rapidly and because the opportunities for the spreading of disease are so much greater

The scope of community health

Social medicine has two major functions preventing disease and, if that fails, curing it The second includes the building of hospitals and clinics, the training of nurses and doctors and the financial support of sick people and their families. This aspect of social medicine is relatively expensive, both because of the cost of the facilities needed and because of the loss of productive work. The preventive aspect of social medicine is less dramatic and goes largely unnoticed, yet it affects all of us

On one level these measures range from the proper disposal of the dead to the myriad regulations, typical of almost all countries today, that control building standards, density of population in towns and cities, the numbers of people that can be carried on ferries or seated in cinemas, conditions for workers in factories, and permissible levels of aeroplane noise, emissions from car exhausts and industrial toxic wastes. The tragedy during the 1950s and 1960s in Minamata Japan, where unsupervised dumping of paper mill effluent into the sea caused the death or disablement through methyl mercury poisoning of those who ate fish or shellfish caught in the area, highlights the need for community action to ensure public health

The same vigilance is needed in testing new drugs or similar products which are brought on to the market before they are thoroughly tested, sometimes with unpredictable and tragic side effects

A clean water supply

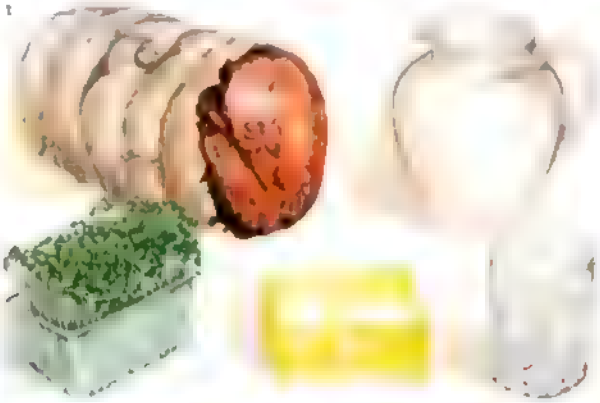
The giant Roman aqueducts that are still scattered across Europe [6] and the ingenious system of underground cisterns or *qanaats* that honeycombed ancient Persia, attest to the vital importance that every community has attached to water And the linked needs of providing houses with a regular supply of pure drinking water and taking away sewage in such a way that the two do not mix are still major preoccupations of health authorities For the contamination of water with even tiny amounts of faeces can cause an epidemic of dysentery, cholera or typhoid

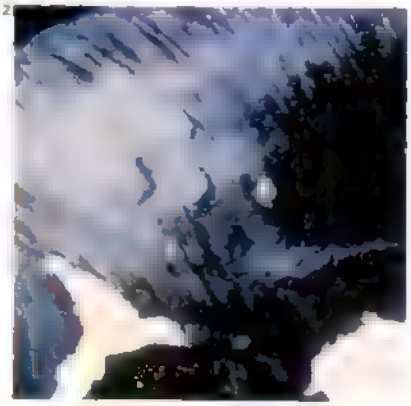
The risk is greatest in developing countries where these diseases are endemic (being highest when an earthquake, typhoon, flood or other disaster disrupts the normal facilities), but can easily occur even in a country as proud of its cleanliness as Switzerland An outbreak of typhoid in Zermatt in 1963, which affected 313 people and killed

CONNECTIONS

See also

1 The risk of infection from contaminated food is high and so all stages of its production, storage, processing and preparation are carefully regulated and monitored Imported food is subject to the same standards and is rigorously checked both to ensure adequate hygiene and to exclude any new animal and crop pests or diseases that are not already endemic in the country

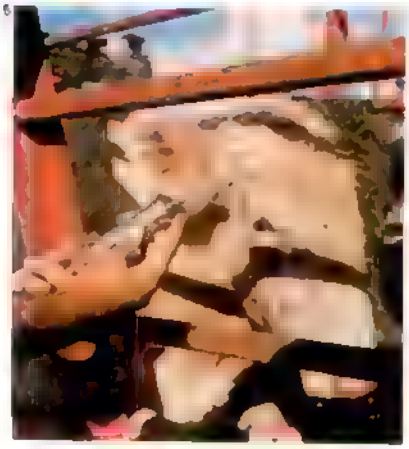




2 Tuberculosis of the bone and brain has been largely eradicated in developed countries over the last 40 years by stringent control of milking herds Cows are tested every year for the presence of tuberculosis bacilli by injecting them with tuberculin A positive reaction (swelling under the skin as shown here) indicates they carry the disease and can pass it in their milk to humans

3 Efforts by restaurants and hotel managements to maintain standards of cleanliness, proper water supply and waste disposal, adequate toilet facilities and staff health are checked by official inspectors.

4 Animals have to be slaughtered in accredited abattoirs both to ensure painless death and to enable the meat to be inspected for the presence of such infections as tapeworms and tuberculosis



5 Vaccination of cattle against contagious abortion or brucellosis not only prevents the loss of calves but also protects humans from the unpleasant undulant or Malta fever found in tainted milk

6 The Pont du Gard, near Nimes, France is one of the best known of the Roman aqueducts still standing. The longest built was the 92km (57-mile) Acqua Marcia from the upper Aniene valley to Rome, of which only 11km (7 miles) was above ground

7 Even fresh vegetables can be a potential source of disease Watercress, for example, can carry on its leaves the eggs of the sheep liver fluke (*Fasciola hepatica*), a parasitic flatworm which causes liver "rot Parts of its life cycle is spent in the fresh water snail so every effort is made to prevent infected snails from entering the beds Water in well-managed watercress farms is drawn from wells or springs that carry no risk of infection and sheep are kept away

three before it was finally eradicated, was traced to faulty sewage disposal

Water has to be properly collected – from springs, rivers or deep bore-holes – stored purified (usually with chlorine, although ozone is sometimes used) and then distributed. In some parts of the world fluoride is added in minute quantities (1 or 2 parts per million) because experiments have shown that its presence decreases the incidence of tooth decay Constant tests are carried out at all stages to make sure that no contamination from sewage or other impurities takes place

The technology exists to clean sewage so thoroughly (and in the process remove the river, lake and sea pollution that is such a feature of industrial countries) that the water can be re-used directly, but the costs involved in these processes are high and, all too often the priority is regrettably low There are usually two or sometimes three stages in the cleaning of sewage, and the first stage alone removes about 30 per cent of organic wastes

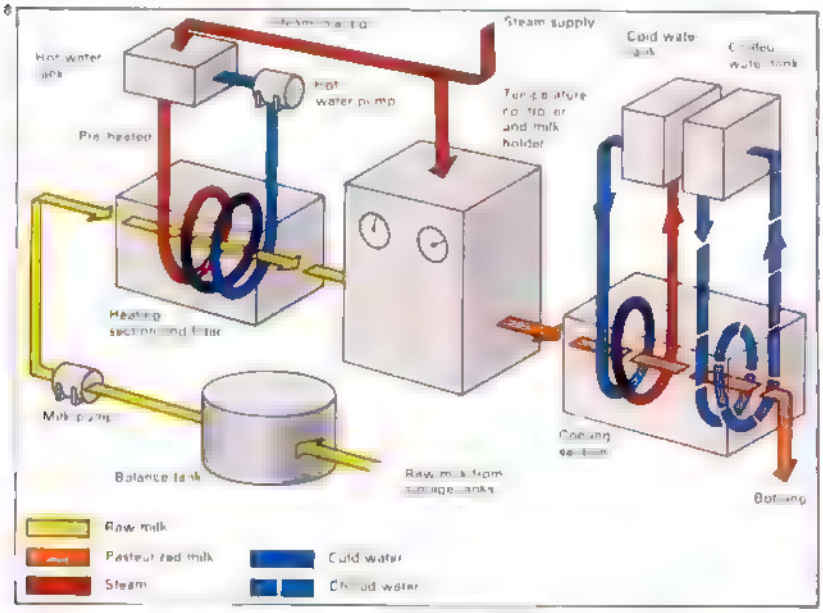
A further advantage of an adequate water supply is that it makes a water-carried sewage system possible. Such a system, if linked by proper sewers [10] to sewage treatment works, greatly reduces the danger of diseases spreading, particularly those such as food poisoning, poliomyelitis and dysentery that are transmitted from faeces to food by flies

Maintaining food standards

All stages of the production of food, from growing and processing to preparation and eating, must be carefully regulated. Techniques for the improvement of food include the addition of vitamins to such common foods as margarine and bread, the pasteurization of milk [8] to prevent the spread of tuberculosis and brucellosis, the inspection of meat to exclude tapeworm infestation [4] and the addition of iodine to salt to combat goitre

But no matter how high the quality of food it can still become a danger to health if it is not properly handled Thus there are laws to ensure hygienic conditions in food factories (faulty canning is particularly hazardous, carrying with it the risk of botulin poisoning), food shops and restaurants [3] The aim is to prevent food becoming contaminated in any way

Public wash-houses were common in European cities up to the 1930s Erected by the authorities they were heralded as a great advance in hygiene and even of substantial moral benefit to the 'poorer' classes Those built in villages simply provided cold running water and rubbing boards, but later town wash houses were equipped with large boilers for hot water and even steam driven spin dryers. Rising wage levels and the advent of cheap electrical washing machines largely removed the need for them and they have been replaced by coin operated laundries equipped with washing machines, spin dryers and often dry cleaning machines



8 Pasteurization of milk destroys the bacteria that cause tuberculosis and undulant fever In the process named after the French scientist Louis Pasteur (1822 95) the milk is heated to 62 71°C (144–160°F), which is lethal to most bacteria but not high enough to alter the quality of the milk As fewer cows carry infectious diseases pasteurization s main use s in improving the keeping qualities of milk

9 Unwanted solid refuse can be disposed of by dumping on land or at sea and by incineration Refuse poses wide environmental problems through its increasing volume and the use of materials that break down slowly

10 The provision of adequate facilities for disposing of storm water and sewage in towns is a major health responsibility of local government Drains and sewers remove such waste for treatment and purification before it is discharged into rivers, lakes or the sea

11 Many diseases are spread by infected animals or their parasites Rabies is mainly transmitted by foxes, dogs, badgers and cats Where there are no barriers to the movement of animals such diseases spread rapidly The sea is the most effective barrier Provided all imported animals are kept in quarantine for the incubation period of the organism, any island, like Britain, can keep such diseases at bay

Preventive medicine

Preventive medicine is the sum of all those measures – general physical check-ups, vaccinations, dental visits and eye tests – that forestall illness either by entirely preventing its occurrence or by catching it so early that it can be easily and quickly treated. Waiting for teeth to ache before visiting the dentist, for example, could well mean extensive filling and even the removal of nerves or teeth.

Care before birth

Understandably, preventive medicine is most in evidence and most widely accepted during pregnancy. Indeed, the process often begins even before conception. At this stage it is often possible to predict the chances of parents with a genetic defect producing a diseased child, and help them to make a decision for or against parenthood. A woman carrying the trait for haemophilia, for example, might decide against bearing a child on being told that any daughters, while directly vulnerable, would have a 50 per cent chance of being carriers and that any sons would have a 50 per cent chance of suffering from the disease.

During pregnancy, an expectant mother makes regular visits to an antenatal clinic so that the progress of the fetus and her own health can be monitored. She is checked to ensure that she is not getting too fat, that blood pressure is not too high, that she is not suffering from diabetes and that she has not contracted syphilis [4] or any of the diseases specific to pregnancy. Particular vigilance is exercised against the metabolic disorder called pregnancy toxaemia.

These antenatal visits also guarantee early action in the event of complications such as narrowness of the birth canal or malposition of the placenta. Monitoring the fetus assures its viability and health and that, at the end of pregnancy, it is in the correct position in the womb for easy delivery.

In recent years a sophisticated new process has been developed to monitor the fetus: amniocentesis [1]. This technique, usually employed when it is suspected that the fetus may have some hereditary defect, involves the passing of a small hollow needle through the mother's abdominal wall into the bag of fluid, the amniotic sac, in which the fetus lies. A little of the fluid is withdrawn and the cells that it contains are grown in tissue culture. These cells are then examined for abnormalities of their chromosomes and chemistry. Not all the abnormalities identified by this method have an ominous significance. However, the discovery of the characteristics of mongolism and Tay-Sach's disease (which causes blindness and severe mental retardation) will force the physician to think of ending the pregnancy prematurely.

Vaccines versus disease

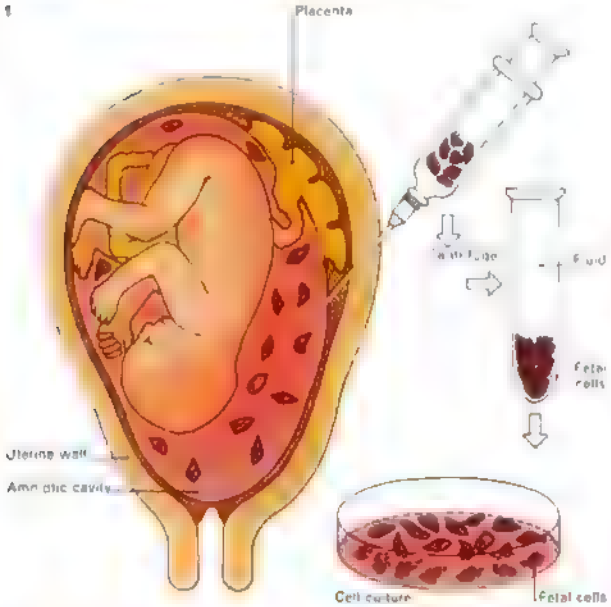
In the early years of life preventive medicine is largely concerned with ensuring proper nutrition and with vaccination against the common infectious diseases of childhood.

The use of vaccines began in 1796 when Edward Jenner (1749–1823) demonstrated the capacity of *Vaccinia* virus to provide protection against smallpox. Today in the temperate countries of the world the objective is to vaccinate all infants during the first year of life against diphtheria, tetanus, whooping cough, poliomyelitis and measles. So effective have these programmes been that, with the exception of whooping cough,

CONNECTIONS

See also

1



1 Amniocentesis the extraction by needle puncture of cells from the fluid around the fetus – greatly increases the chance of identifying the presence of hereditary diseases. At present the fetus's sex and some of 40 of the more than 1 600 genetic diseases can be detected after the cells have been grown in culture. The best time for the "tap" is about the 16th week of pregnancy. Problems of using this technique are its delicacy and the risk of long-term damage to the fetus. The possibility of its being used for eugenics or genetic engineering and antenatal sex determination raises questions about the social effect of wide use.

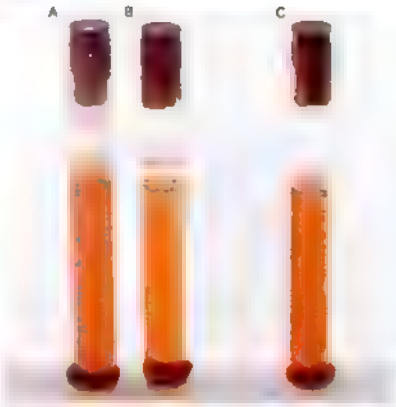
2

2 The fetal heart beat becomes clearly audible through the obstetrician's special stethoscope after 28 weeks. More advanced electronic devices can detect the heartbeat as early as seven weeks.

3

3 Vaccines are produced by artificially cultivating the illness causing viruses or bacteria. They are collected from infected animals or humans and grown in conditions suited to fast development as here. The micro-organisms produced are washed, centrifuged in a sterile salt solution and killed. The viruses are finally weakened to make them less virulent. A dose injected into the body induces the blood to form antibodies, which combat invasion by the disease.

4



4 Maternal syphilis is a serious threat to an unborn child. The blood of pregnant women is always examined for this infection, usually by the Wassermann reaction. Syphilitic blood inhibits the activity of a substance called "complement", which dissolves specially prepared red blood cells of sheep. These test tubes show a positive control (A), a negative control (C) and a positive test (B). Fever-causing diseases such as malaria can produce a false positive reaction.

all these diseases are increasingly rare Tuberculosis has presented more enduring problems, but mass miniature radiography [6, 9] for early diagnosis and the use of antibiotics for treatment have reduced the threat of this once epidemic disease to almost negligible proportions.

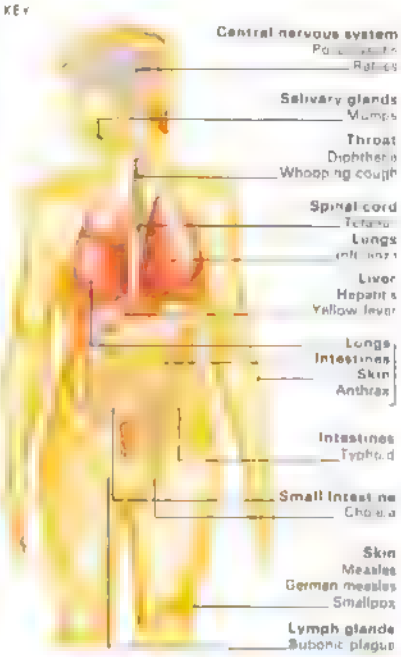
In some circumstances extra vaccinations are provided on a more selective basis German measles (rubella) vaccine is used for adolescent girls who have not had the disease, to prevent the risk of an attack during pregnancy There are also influenza vaccines, rabies vaccine for the staff of quarantine centres and zoos, anthrax vaccine for veterinary surgeons and hide porters, and a vaccine for hospital workers likely to contract hepatitis as a consequence of handling blood. The length of time for which these vaccines give protection varies For cholera it is six months, for yellow fever a lifetime

Preventive medicine for adults

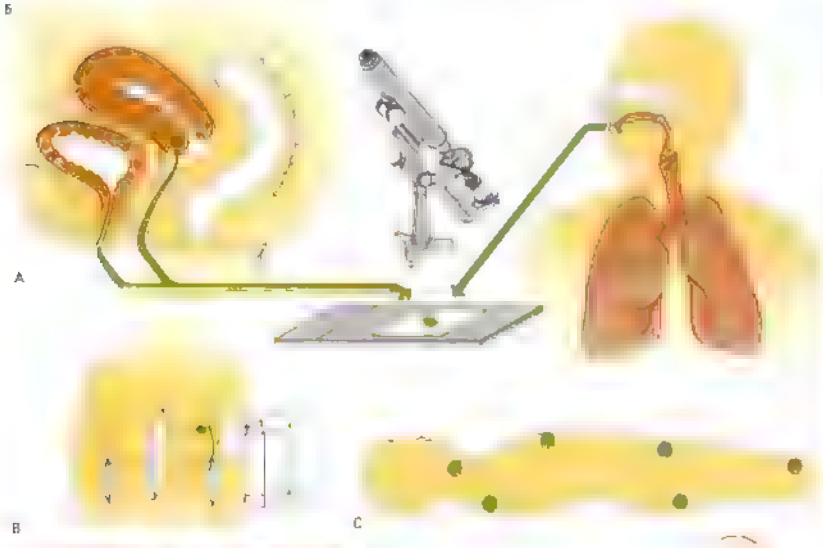
In later years the emphasis in preventive medicine changes from absolute prevention of infectious illness to early diagnosis by periodical check-up of the illnesses that threaten in middle life

A physical examination combined with measurement of blood pressure, an X-ray of the chest, an electrocardiograph and an examination of the urine for the tell-tale sugar of diabetes would constitute a typical thorough check on the more common diseases and ailments of middle age [8] In more elaborate examinations the blood is tested for excess fats and cholesterol

For women, there are extra hazards associated with their reproductive systems and special clinics are today becoming more common in developed countries. The functions of such clinics include early diagnosis of cancer of the breast and of the cervix. Careful palpation of each breast usually reveals any lumps suggestive of cancer, but more complex methods using soft X-rays and heat sensors are sometimes used [5]. Examination of the cervix is usually accomplished by the cervical smear method in which the cells from the cervix are smeared on a glass slide and examined microscopically for evidence of cancerous changes [7]

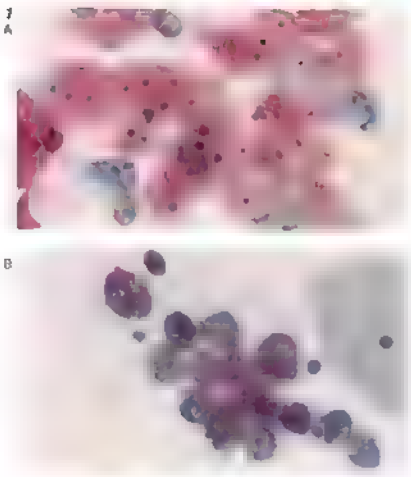


The prevention of infectious diseases through vaccination is an important part of preventive medicine The administration of most vaccines causes the production of substances known as antibodies in the blood of the vaccinated person These, either alone or in concert with other components of the blood plasma or with the white blood cells attack invading microbes and so prevent infection The effects of antibodies are quite specific and those induced by a vaccine intended to prevent one infectious disease give no protection against other diseases Shown here are the various diseases that can at present be combated by vaccines and the sites of their attack

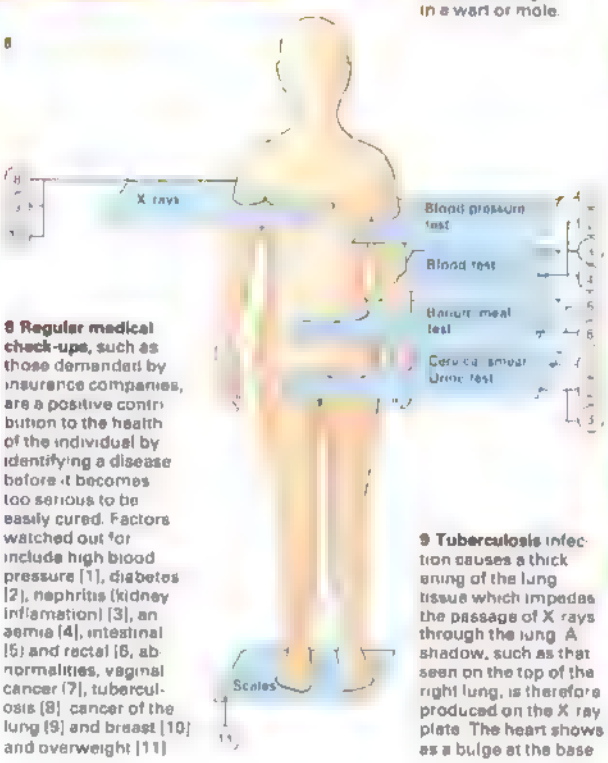


5 Early detection of cancer is the best way of minimizing tissue damage and death Screening tests include microscopic examination of cells (A) and mammography (B), an X ray technique used to detect breast lumps and to determine whether they are malignant This crucial early detection is aided by individual alertness to certain danger signals (C) including hoarseness or a nagging cough difficulty in swallowing, breast lumps, change in bladder habits, unusual bleeding or discharge, a persistent sore and an obvious change in a wart or mole.

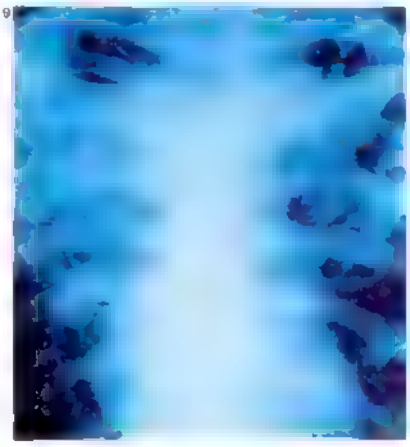
6 Mass miniature radiography is a cheap way of examining the lungs by X ray and has been instrumental in the campaign to eradicate tuberculosis. X-rays passing through the patient's chest fall on a fluorescent screen to form a picture similar to that on an X-ray plate This picture is photographed and the small photographic negative can then be examined by the radiologist



7 Microscopic examination of cells can help detect several different cancers A smear taken from the cervix during what should be a yearly examination would reveal whether the cells are normal (A) or have undergone cancerous changes (B) Cells from the urinary tract shed with urine may indicate kidney, ureter or bladder cancer Lung cancer may be verified by examination of sputum, while oral cancer can often be detected by examining cells scraped from the mouth



8 Regular medical check-ups, such as those demanded by insurance companies, are a positive contribution to the health of the individual by identifying a disease before it becomes too serious to be easily cured. Factors watched out for include high blood pressure [1], diabetes [2], nephritis (kidney inflamation) [3], anemia [4], intestinal [5] and rectal [6] abnormalities, vaginal cancer [7], tuberculosis [8] cancer of the lung [9] and breast [10] and overweight [11]



9 Tuberculosis infection causes a thickening of the lung tissue which impedes the passage of X rays through the lung A shadow, such as that seen on the top of the right lung, is therefore produced on the X ray plate The heart shows as a bulge at the base

Natural remedies

As civilization developed, the stock of folk lore about the curative properties of various substances or rituals was gradually classified. This was the first step towards scientific medicine. Once alleged remedies were listed and classified those that stood the test of time could be noted. However, it was not until the nineteenth century that major advances were made in understanding why some plant extracts, for example, had curative properties, for by then chemistry had advanced to the stage at which the molecular structures of individual substances could be identified, so that they could be synthesized in the laboratory and tested therapeutically.

Early natural remedies

Among the earliest pharmacopoeias (catalogues of medicines) was that compiled by Dioscorides (*fl.c.* AD 60), a surgeon in Nero's army. During his travels he made notes about the remedies used in different countries. These notes provided the basis for a tradition that, although largely lost after the fall of the Roman Empire, passed back to Europe from the Arabs in the fifteenth century. At this time pharmacists were known as *aromatari* because the materials they sold were aromatic substances extracted from plants. Remedies with an animal or mineral base were also used. Paracelsus (*c.* 1493–1541) was as renowned for his cures with mercury as for those that used opium.

At this time there was a continuous interplay of different factors that advanced the science of pharmacy on several fronts simultaneously. Although there were several thousand known "remedies" for various ailments one school of medical theorists still argued that there was, albeit undiscovered, a substance that would cure all ills. This idea of the "cure-all" is in the alchemical tradition of a philosopher's stone that will turn other materials into gold and the alkahest, or universal solvent, that will dissolve all other substances. In their search for such universal substances the alchemists and their followers stumbled on some useful processes. Raymond Lully (*c.* 1235–1315) is credited with discovering how to prepare pure alcohol, he probably discovered how to react this with acid to produce anaesthetic ether.

The availability of pure alcohol was a great advantage in the preparation of tinctures and essences because many of the active principles in medicinal plants dissolved more readily in this solvent than in water. And as the developed European countries discovered other parts of the world, a great variety of new plants became known. These were added to the pharmacopoeias and mixed with other substances in alcoholic solutions.

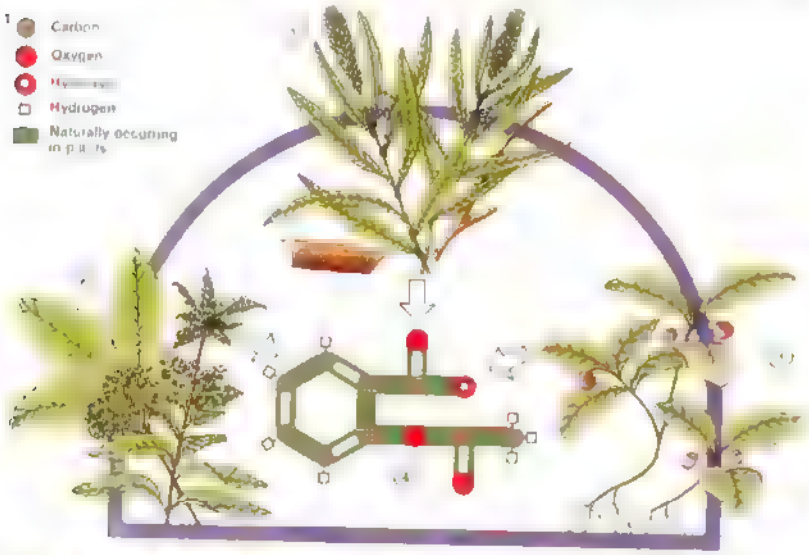
Isolating the active ingredients

The major problem was to sort out which of the substances were effective and which could be left out. In the eighteenth century when the physician William Withering (1741–99) was introduced to an old country woman whose secret herbal mixture seemed to be surprisingly effective in cases of heart failure, he had to sort through 20 different ingredients to discover that it was the foxglove leaves that were effective. (The fact that they contained the substance digitalis was not discovered until later.)

By the time the New and Old Worlds had been thoroughly explored the number of

CONNECTIONS

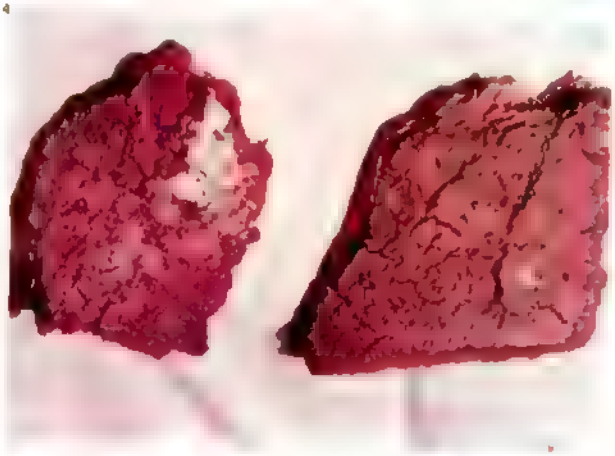
See also



1 Aspirin the most widely consumed drug in the world, is a chemical substance called acetylsalicylic acid. Plants such as willow (*Salix fragilis*) [1], meadow sweet (*Filipendula ulmaria*) [2] and shallon (*Gaultheria shallon*) [3] contain the related substance salicin. The presence of this pain reliever was the basis of many old remedies. As salicin is too bitter to be taken internally (the same is true of salicylic acid, the first chemical derivative of salicin) it was not until it was made into its acetyl derivative in 1899 that aspirin [4] emerged.

2 Aspirin reduces fever [A] by regulating the brain's temperature control centre [1] and by increasing sweating [2]. It reduces pain by cutting reverberatory impulses at the thalamus [3] and of inflammation at joints [4, 5], but it can cause bleeding [6] and skin rashes [7] with some people. Aspirin reduces inflammation [C] by limiting production of the chemicals prostaglandins [illegible] which are [illegible] in inflammatory cascade [B] when the body's defences track e an antigen or foreign body. Histamine is not affected.

3 From the roots of *Rauwolfia serpentina*, reserpine, a valuable tranquillizer is obtained. Analysis of its chemical structure has led to the development of more effective synthetic tranquillizers with no damaging side-effects.



4 Quinine is a bitter substance derived from cinchona bark. Brought from Brazil by the Spaniards, it is a remedy for malaria and has only recently been superseded by synthetic compounds.

natural plant substances with known pharmacological action was enormous. There was opium from poppies [8], emetine (a remedy for amoebic dysentery) from the ipecac plant, quinine from cinchona bark [4], castor oil from castor seeds and salicin from willow [1]. A number of other plant substances with major effects on living organisms were also known including belladonna, strychnine and curare [7]. All were used as poisons but are now valuable in medicine. But the physiological effects of these substances are so powerful that they must be used in controlled doses. It was only as the active ingredients were isolated that it became possible to use them quantitatively.

Sophisticated techniques

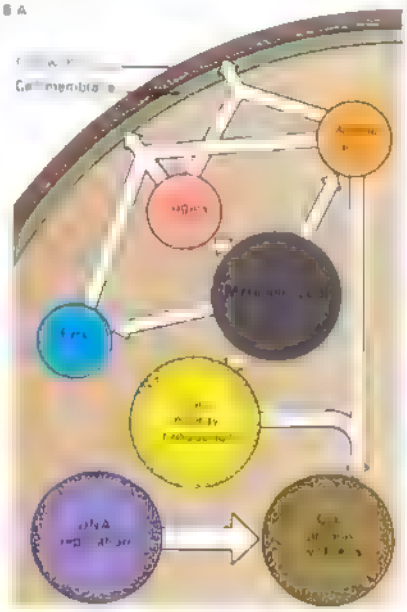
During the nineteenth century chemistry developed rapidly as an exact science and many plant substances were isolated as pure crystalline compounds. They were all classed as "organic chemicals" and it was not until 1828 that scientists realized it was possible to make such chemicals in a laboratory. In the same year Friedrich Wöhler (1800–82) synthesized urea from ammonia cyanate and changed the entire emphasis of organic chemistry. Until that time it had been believed that such substances could be produced only by living processes. Nevertheless the chemical structures of many of the pharmacologically useful plant substances are so complex that it is more economical even now to extract them from plants than to make them in the laboratory.

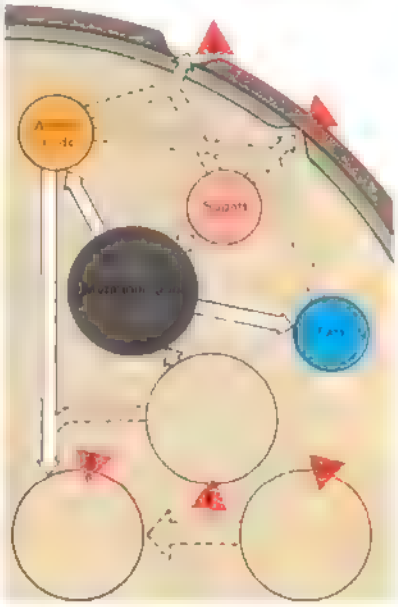
Just how complex some of these substances are was not discovered until well into the twentieth century when sophisticated techniques of molecular analysis were devised. But it is not necessary to understand the structure of a chemical completely in order to modify it chemically. This frequently produces a related substance that will share some properties of the original substance but not others. By trial and error chemists gradually began to improve upon natural remedies producing analogues of natural compounds that were safer, more effective or had less unpleasant side effects. Thus began today's pharmaceutical industry which not only modifies nature, but has added much of its own.

Nearly 2,000 years ago a brew made from white willow leaves was recommended for gout. To kill pain today we take an aspirin. Now solely a product of the chemical laboratory, aspirin can also be made from willow brew.

5 Mould on bread is usually viewed with disgust, but it is from variants of one common bread mould that penicillin is derived. This is doubtless why "mould poultices" are an old effective remedy.

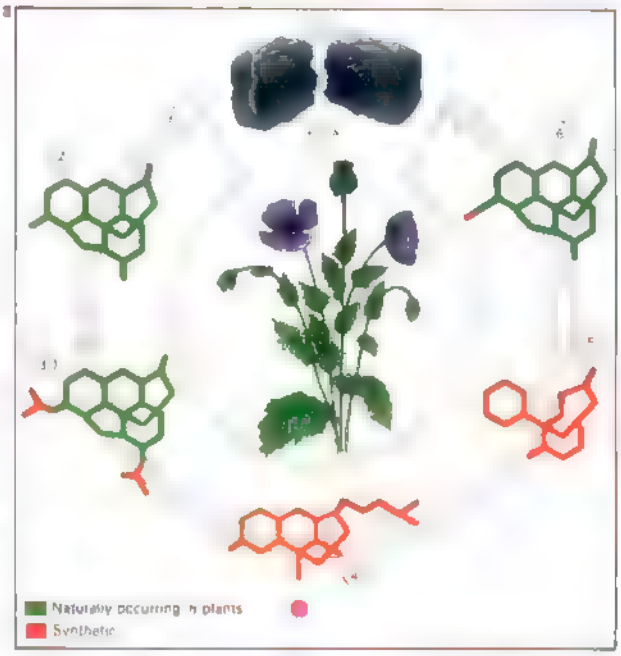
6 Penicillin, which is derived from fungi, is now only one of many antibiotics that kill or inhibit the growth of other micro-organisms. This makes them useful in treating bacterial infections. The normal bacterial cell processes [A] are attacked by different antibiotics [B] which may alter cell membrane structure [1] or inhibit cell wall synthesis [2], protein synthesis [3], energy production [4] or DNA replication [5].





7 South American Indians carry pots of raw curare into which they dip their arrows to render them poisonous. This toxic substance is used in controlled doses in surgery for total muscular relaxation.

8 Most derivatives of opium [1] such as morphine [2], heroin [3] and meperidine hydrochloride [5] although extremely effective pain-killing (analgesic) drugs are at the same time highly addictive. It is possible in the laboratory to pinpoint which parts of the molecule are responsible for the analgesic and which for the addictive effect. On this basis drugs such as codeine [6] and pentazocine [4] can be designed and synthesized. Extensive clinical trials are necessary to confirm their safety.



Man-made cures

At the beginning of the twentieth century most effective remedies were still derived from plants. Now they are produced in factories following directions worked out in chemical laboratories.

The origins of the drug industry

In the late nineteenth century the nature of chemical structures came to be understood and research chemists, notably in Germany, synthesized hundreds of new organic molecular compounds based on carbon atoms. As a result the idea of a chemically based pharmaceutical industry emerged. Attempts to make synthetic versions of natural compounds on the basis of incomplete knowledge often led to curious results. The first synthetic dyestuff, mauveine, was produced accidentally in 1856 by William Perkin (1838–1907), while he was trying to make quinine.

The result of Perkin's discovery was the start of the synthetic dyestuffs industry. Within a few decades hundreds of synthetic dyestuffs had been manufactured. Ironically, it was from this wealth of new material that chemotherapy, the modern concept of treating disease with chemicals, came into being. The German scientist Paul Ehrlich (1854–1915) [2A], discovered that certain dyestuffs stained only specific tissues when he treated microscopic specimens with them. From this he conceived the idea that some dyestuffs might selectively and safely destroy the micro-organisms that cause diseases.

Ehrlich tested 500 different dyestuffs on mice that had been infected with trypanosomes – blood parasites that cause sleeping sickness and other diseases. None of them worked. He then tested other compounds with structures similar to the dyestuffs but which also contained atoms of arsenic. His compound "606", or Salvarsan, although ineffective against trypanosomes, turned out to be effective against the bacteria that cause syphilis. It was first manufactured in 1910.

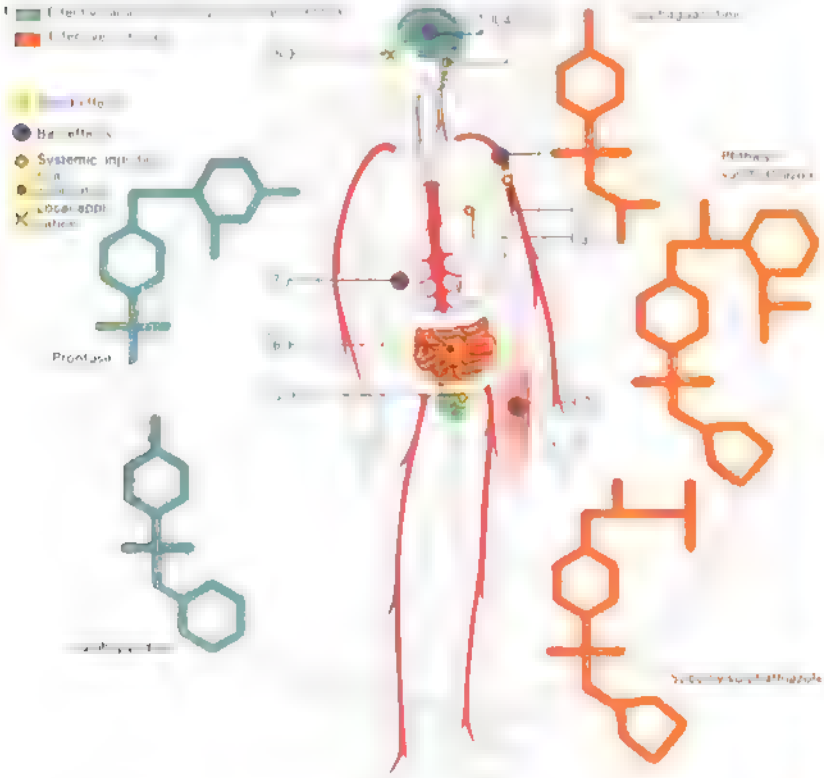
The sulpha drugs and penicillin

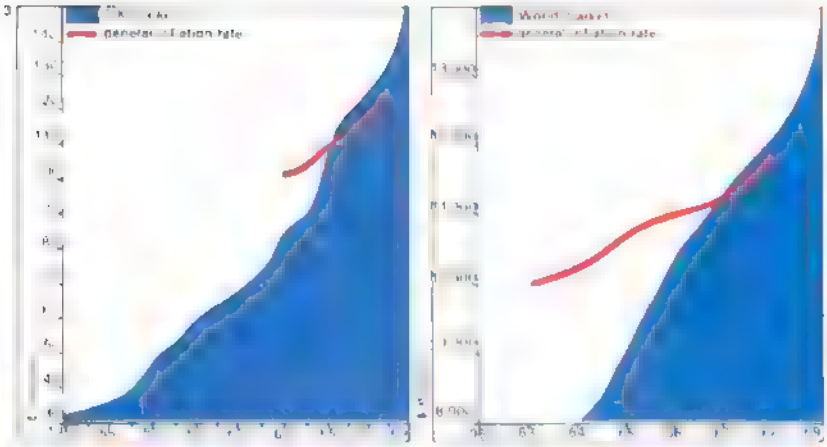
Another German scientist, Gerhard Domagk (1895–1964), was responsible for the second great therapeutic discovery to come from the dyestuffs industry. In the early 1930s he discovered that the red dye Prontosil was an effective bacteriostatic – a synthetic substance that prevents the proliferation of bacteria in the bloodstream. Until that time the most dangerous part of many surgical operations had been the period after the operation when the patient was liable to die from septicaemia – infection from bacteria introduced into the body during the operation.

French scientists soon discovered that only one part of the Prontosil molecule – sulphanilamide – was effective. By the end of the 1930s the sulpha drugs [1] had lowered post-operative mortality dramatically and, in 1938, a bonus appeared in the form of a modified sulphanilamide molecule – sulphapyridine – effective against tuberculosis.

At about the same time the bacteriostatic properties of the penicillin mould, discovered by Alexander Fleming (1881-1955) [2B] in 1928, were being further investigated by Howard Florey (1898-1968) and Ernst Chain (1906–) [2C]. The development of penicillin manufacture during World War II heralded the age of antibiotics. Many micro-organisms were found to contain complex

See also





1 The sulpha drugs act by preventing the multiplication of bacteria and so give the body's natural defences the chance to overcome them. The drugs combat blood [1] and urinary tract [2] infections, pneumonia [3], meningitis [4], eye infections [5] and alimentary tract [6] infections such as bacillary dysentery. Over doses [7] can produce skin rashes, raised temperature, impaired kidney function and anaemia.

2 Paul Ehrlich [A] is regarded as the founder of modern chemotherapy. He believed that individual chemicals could act as 'magic bullets' against the agents of infection and his discovery of Salvarsan proved his case. A second era of expansion for the drug industry was based on the accidental discovery by Alexander Fleming [B] that moulds could produce antibacterial substances. Ernst Chain [C] who helped in the development of penicillin, took chemotherapy a stage further by suggesting that semi-synthetic drugs would prove useful.

3 The growth of the pharmaceutical industry during the twentieth century has been staggering. As medical care has spread so the demand for new and better drugs has increased. This demand can only be met by extensive research programmes. Shown are the rising values of drug sales during the 1960s in the UK and the world.

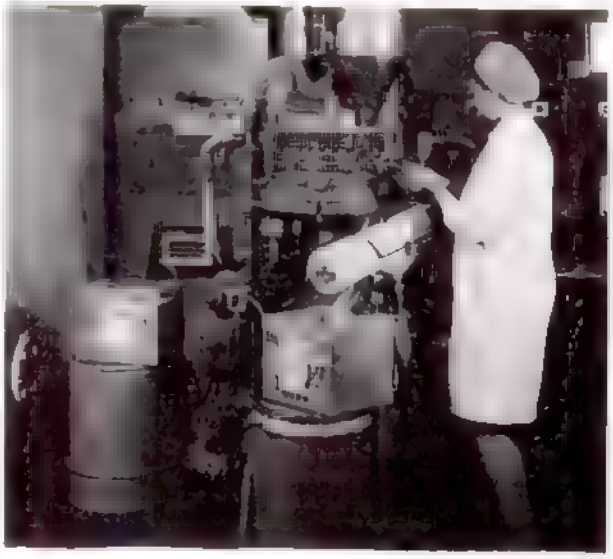
chemicals that would combat other micro organisms. Streptomycin was discovered in 1943 and since then dozens of new antibiotics have been discovered and manufactured on a commercial scale [Key, 3]

In most cases it is cheaper to cultivate the micro-organisms by fermentation and then extract the active materials, than to synthesize these materials in a laboratory. A compromise between the two approaches was the development of semi-synthetic penicillins in the 1960s. In 1959 the key fragment of the penicillin molecule, 6-aminopenicillanic acid, was isolated. When this is reacted with various organic molecules it is possible to manufacture penicillins that do not occur in nature. These tend to attack micro-organisms that have developed resistance to natural penicillins, and they are therefore of greater use therapeutically [6]

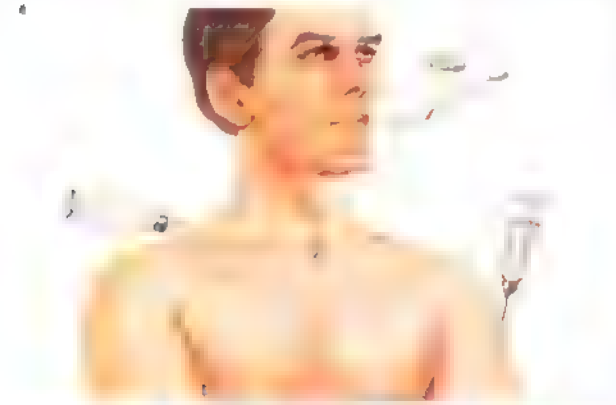
Natural and synthetic products

The concept of modifying natural molecules has spread dramatically through the pharmaceuticals industry. The contraceptive pill, developed in the early 1960s, emerged as the result of the chemical modification of natural sex hormones, which had been isolated in the 1930s. The chemical structures found in human sexual and adrenal cortical hormones are also found in slightly different forms in many living organisms. By chemical modification of a starting material obtained from wild yams it has been possible to manufacture the large quantities of oral contraceptives now in demand.

Many other natural products, such as vitamins and adrenaline, have been synthesized in laboratories and some of them are made commercially by synthetic methods. Adrenaline belongs to an important class of chemical compounds called phenthylamines, this also includes ephedrine, an old Chinese plant remedy now used in asthma treatment and in "pep pills". Many completely synthetic drugs such as barbiturates have been discovered by accident. The results of unpredicted side-effects, such as the impairment of limb development caused by thalidomide, mean that rigorous safety checks have to be made and many governments have set up safety committees to test new drugs [5]



The high-speed tableting machine has supplanted the apothecary's crude apparatus. This is a natural consequence of the need for large amounts of drugs and strict quality control



4 Modified forms of penicillin mean that the drug can be given orally, by injection or by local surface application. Many different sorts of penicillin are now available, all with slightly different molecular structures. These slight differences can mean quite large differences in effect. Apart from self administered substances such as aspirin, penicillin is probably the most commonly used antibiotic the world over

5 Samples from each batch of every manufactured drug must be sent to analytical laboratories. There checks can be made to ensure that consistent standards of purity are always maintained

6 Antibiotics have been increasingly used in recent years to keep animals healthy and to combat disease in humans. A disturbing side-effect has been that the bacteria they used to kill are becoming immune

7 Viruses are resistant to most synthetic drugs. To find a cure for viral diseases such as the common cold scientists experiment, after artificially infecting a volunteer, with complex naturally occurring defence substances

8 Drugs can be formulated for oral consumption as pills, powders, solutions or suspensions. Different colours and shapes are used to distinguish the thousands of products now marketed

Radiology and radiotherapy

The effects of radiation may be used both for investigating diseases, in diagnostic radiology, and for treating them, in radiotherapy. Much of the early work on the phenomenon of radioactivity was undertaken by Marie Curie (1867–1934) [Key] who, with her husband Pierre (1859–1906) and jointly with Antoine Becquerel (1852–1908), was awarded the Nobel prize for physics in 1903. In 1911 she was also awarded the Nobel prize for chemistry for the discovery of radium and polonium. Her death from leukaemia was probably due to over-exposure to radiation.

Over-exposure illnesses and the long-term effects of atomic bombs point very clearly to the powerful and dangerous effects of radiation waves. By careful and controlled use, however, many of the properties of beta, gamma and X-rays are exploited in medicine for both diagnosis and treatment.

X-ray photography

Most kinds of energy waves are able to pass through some materials and not others. Light waves, for instance, pass through air and also through some liquids (such as water) and even solids (such as glass), but not through walls and doors. X-rays on the other hand can pass with varying ability through body tissues, but can be blocked and absorbed by other substances, such as lead, which is often used in radiology. In the same way that varying amounts of light waves can produce a picture on a photographic plate, so also can X-rays activate such a film [1].

A simple X-ray machine is rather like a camera. An X-ray source is beamed at the part to be examined and focused on to a photographic film, where the image resembles a film negative. An X-ray picture of a hand, for example, shows bones (which stop X-rays quite effectively) as white, the other tissues as greys and the area around the hand (which receives full X-ray exposure) as black. In order to show up other organs within the body more clearly, radio-opaque substances (ones that do not allow X-rays through) are used. Barium sulphate is one such substance and is used in barium meals and enemas to show up the upper and lower bowel respectively [3]. Other substances injected or given by mouth, often containing iodine, are particularly concentrated in certain body organs. This enables better pictures to be obtained of kidneys and bladder by monitoring the appearance of the dye.

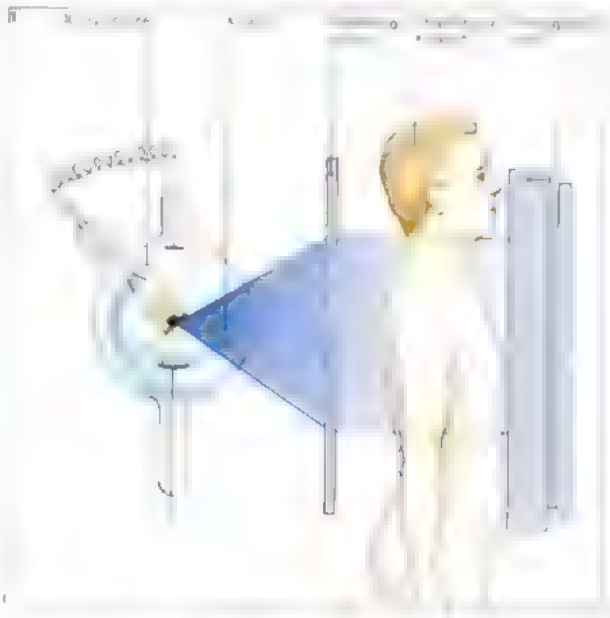
Scanning techniques

Pictures obtained by conventional X-ray machines are not as clear as ordinary photographs. Organs within the body are seen superimposed upon each other and the picture is a composite one. To see an organ or part within the body with less overlaying detail the technique of tomography is used [2]. Here the X-ray source, and the detection system or photographic plate, are rotated in an arc round the body with the specific organ as the centre of the arc. This organ then appears relatively still and in focus, whereas other parts are less clear.

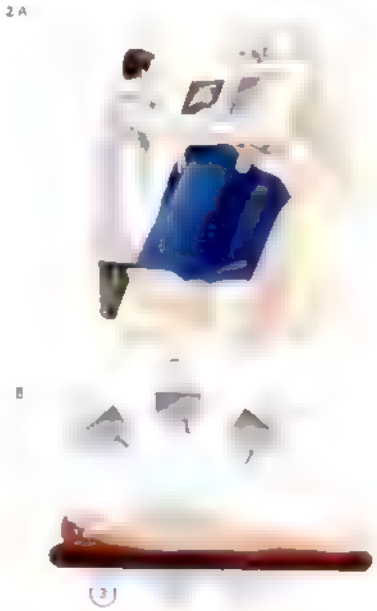
In a major advance, this principle has been used in a very sophisticated manner to scan the brain and, using a computer, to integrate the information. The EMI scanner [5] produces a picture of a "slice" through the brain, and can produce a series of such pictures to give a view of the whole brain. By the use of a

CONNECTIONS

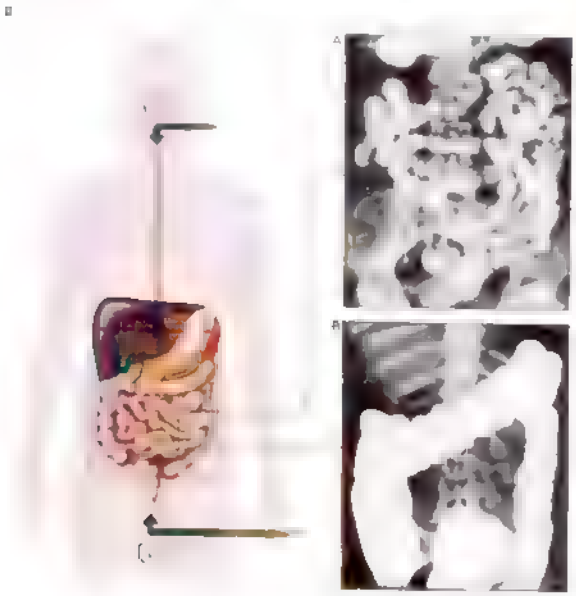
See also



1 The use of X-rays in medicine to photograph the body internally has enormously advanced the diagnosis of disease. Since X-ray waves emerge from an X-ray source travelling in several different planes, they are first passed through a polarizing screen. This organizes or directs them into a single plane and so concentrates them. The body's various organs and tissues (skin, bone, brain, lungs, liver, arteries and so on) absorb varying amounts of the X-rays passing through them. Hence the X-rays striking the photographic plate are of varying degrees of intensity. The result is an image that resembles a film negative.



2 Radiology uses X-rays to visualize internal parts of the body as an aid to diagnosis. Various machines are used to produce X-rays of controlled energy levels and to focus the rays and control exposure. A chest X-ray [A] can be viewed on a television screen, stored on videotape or made into X-ray films. To facilitate diagnosis, substances that do not transmit X-rays can be swallowed or injected to provide a means of contrast. An X-ray film is a negative and the contrast appears white. By rotating the X-ray source [B] and taking pictures from several angles [1, 2, 3] it is possible to focus on organs in a particular plane of the body.



3 Barium sulphate is used in radiology to provide contrast in X-ray photography. Internal tissues such as the digestive system and kidneys allow the passage of X-rays to varying extents. But when filled with a substance that blocks the X-rays, they can be made opaque and show up as white. The barium meal, a harmless paste that passes through and out of the intestines unchanged, can be swallowed [1] or injected into the rectum [2]. When swallowed, it outlines in turn the oesophagus or gullet, the stomach and the small intestine [A]. Ulcers, tumours and alterations of function can be seen. Similarly, the colon or large bowel can be photographed [B].

4 Arteriography is a specialized technique of radiology developed to show up the inside of arteries so that thickening and furring can be detected. This tissue damage in arteries is the degenerative disease known as arteriosclerosis. It contributes to coronary and strokes and to poor circulation in the legs. The arteries are shown up by the injection of a radio opaque dye which reveals irregularities of the vessel wall. The picture here, obtained after injection of a dye into the aorta, shows the large arteries of the pelvis and upper legs and a local narrowing of the artery (arrowed) that is interfering with the blood flow.

computer, a very much more detailed and precise picture is obtained. This technique is now likely to replace many of the techniques in which dyes or air are injected as contrasts into the fluids bathing the brain, for these latter procedures are longer, more complicated and potentially dangerous. A computer scanning system is more expensive (about US$320,000) than those with more conventional machines, but the X-ray dose for the whole procedure is smaller, comparable to that of a skull X-ray. The EMI Whole-Body Scanner, which produces a scan of a body section that is processed in 20 seconds, is now also in use.

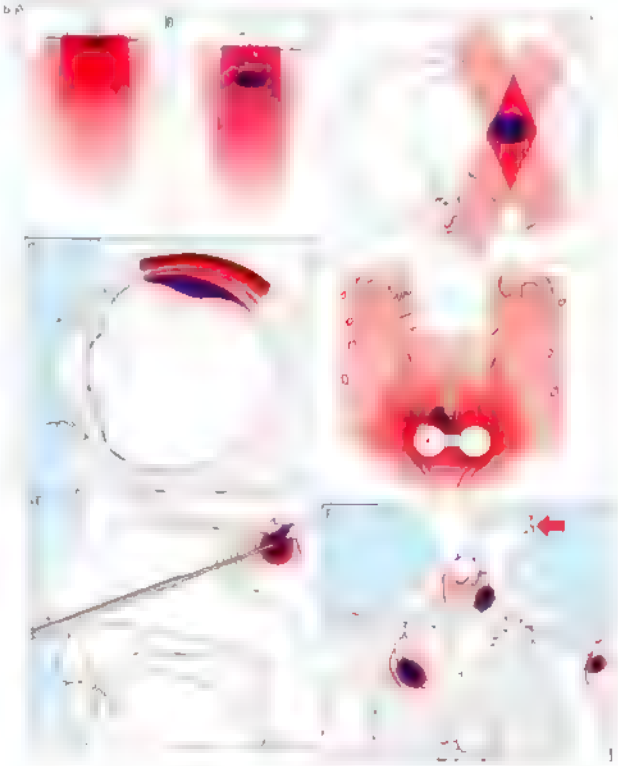
X-rays are not the only type of radiation that can be used to obtain information about the body. Gamma waves emitted from certain radioactive isotopes (forms of certain elements) are also able to pass through body tissues. Small doses of such isotopes can be injected and a measure of the radioactivity taken up by a particular organ may give evidence about it. The thyroid gland, for example, is responsible for using nearly all the iodine taken into the body. If a radioactive iodine isotope is injected into the blood stream, the activity of the thyroid can be gauged by measuring its rate of uptake of iodine. This can be done by measuring the gamma waves emitted from the region of the thyroid. In addition, scanning the thyroid will give a picture of the activity of each part of the gland. Parts that are over- or under-active may thus be detected.

Treatment of cancer by radiotherapy

Using high doses of isotopes can cause tissue damage. In treating thyroid cancer doses several thousand times higher than for diagnosis are used to take advantage of this damaging effect to kill the cancer cells. This is the principle of radiotherapy, where high doses of radiation are beamed on to tumours to kill them [6]. All living cells are generally sensitive to the effect of radiation, and cancer cells are particularly susceptible. Care must be taken to judge the dosage required and to apply it to the correct area. Some cancers respond better than others to rays of high energy, and for these radiation of a particular energy may be required.

Marie Curie's work on radium led to X-ray and isotope techniques for both the diagnosis and treatment of disease.

6 Radiotherapy uses radiation waves to treat disease. Ionizing radiations (beta, gamma or X rays) cause changes in cells (particularly when they are dividing) that kill them. This can be used against malignant cells, which are more sensitive to the energy released by irradiation than healthy cells. Treatment of cancer can be by a beam of low-energy rays focused on to the skin [A] or of higher energy [B] focused below the body surface [1], sometimes from more than one direction [2]. A local surface applicator can be used, as in the eye [C] or combined with beam treatment, as in the neck of the womb [D]. Minute radioactive pellets can be placed deep in the body to treat the pituitary [E], or an isotope administered by mouth may be taken up selectively, as is iodine by thyroid cancer even if the cancer has spread [F].



5 The EMI scanner allows X-raying of the brain through an accurately aligned X-ray source with multiple detectors that move past the patient. Information is fed into a computer and enables a complex picture to be built up. The X-ray unit [A] for examining the brain obtains information easily and simply with low X-ray dose and no distress to the patient. The Whole-Body Scanner produces pictures of a slice through the body [B]. Shown is the abdomen with liver (right).

7 To treat a tumour, radiation is delivered by sophisticated machinery with the operators protected by lead screening from the powerful X-ray or gamma ray source. The patient is also screened so that only a specified area of the body is irradiated for the correct time. Treatment in stages minimizes side effects. The type and energy of the rays are vital for success and new, very high energy rays can now treat previously unresponsive cancers.

Development of surgery

Surgery is the manual treatment of injuries, deformities and other disorders. Early man probably practised simple forms, such as setting a fractured limb, binding a wound, lancing an abscess or pulling a tooth, and there is evidence that he may have attempted delicate operations with some success.

Advances in anatomical knowledge

Even 10,000 years ago "medicine men" were capable of performing the difficult operation known as trephining – the removal of a piece of skull to cure headaches or madness – and the Romans removed endangered babies from their mothers' wombs by what came to be known as Caesarean section. But generally surgery was limited to simple external operations and excluded anything internal. The sophistication of modern surgery is really the result of discoveries made in the last 300 years.

A surgeon's first need is a clear and accurate understanding of the working of the human body. Until late in the Middle Ages surgery remained a primitive trade (closely allied to that of the barber) because neither doctors nor surgeons had any real knowledge of anatomy. The only way to find out is by dissection of the dead, a process looked upon with horror by Church and state alike. Only in the fifteenth century was this taboo even partially overcome, but from that time onwards doctors were able at last to begin their investigation of the structures that lie within the body. Ambroise Paré, often called the father of modern surgery, was one of the earliest to use this new knowledge [1]. Important subsequent landmarks were the demonstration by William Harvey (1578–1657) of the circulation of the blood and the anatomical discoveries of William Hunter (1718–83) and his brother John (1728–93).

Eliminating pain and infection

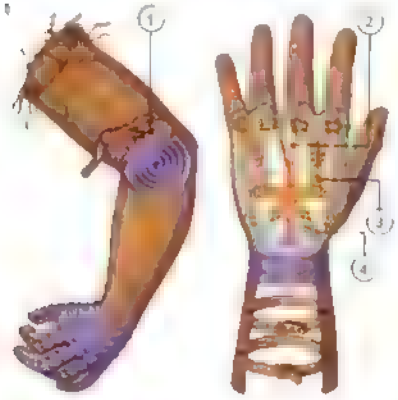
Even the most exact knowledge of anatomy does nothing to solve the two major problems confronting the surgeon and, more important, the patient – namely pain and infection. It was the nineteenth century that brought about the great revolution that transformed surgery from an agonizingly brutal and dangerous business into the relatively safe and painless one that it is today. The first advance came in 1844 when Horace Wells (1815–48), an American dentist, used laughing gas (nitrous oxide) when extracting a tooth. Within a few years ether [2] and chloroform were also being used to eliminate pain during major surgery. William Morton (1819–68) pioneered the use of ether (general anaesthesia).

The second factor in the surgical revolution was the advent of antiseptic surgery. Until mid-Victorian times it was commonplace for patients to develop lethal post-operative infection.

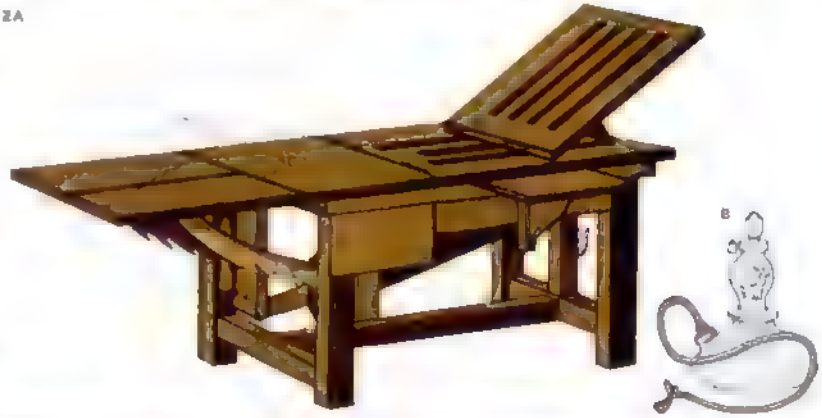
The reason for post-operative infection remained hidden until Louis Pasteur (1822–95) showed that it was caused by germs. The first surgeon to grasp fully the implications of Pasteur's discovery was Joseph Lister, who eventually realized that antiseptic chemicals such as carbolic acid could kill germs and protect operation wounds from infection. His technique of placing surgical instruments in antiseptic solution, of spraying carbolic acid in the air during operations and of applying antiseptic

CONNECTIONS

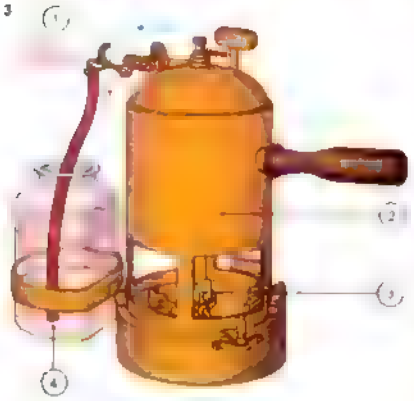
See also



1 French surgeon Ambroise Paré (1517–90) designed this artificial arm. A ratchet and pawl [1] locked the elbow in any position, springs and catches [2, 3, 4] opened and then closed the fingers.

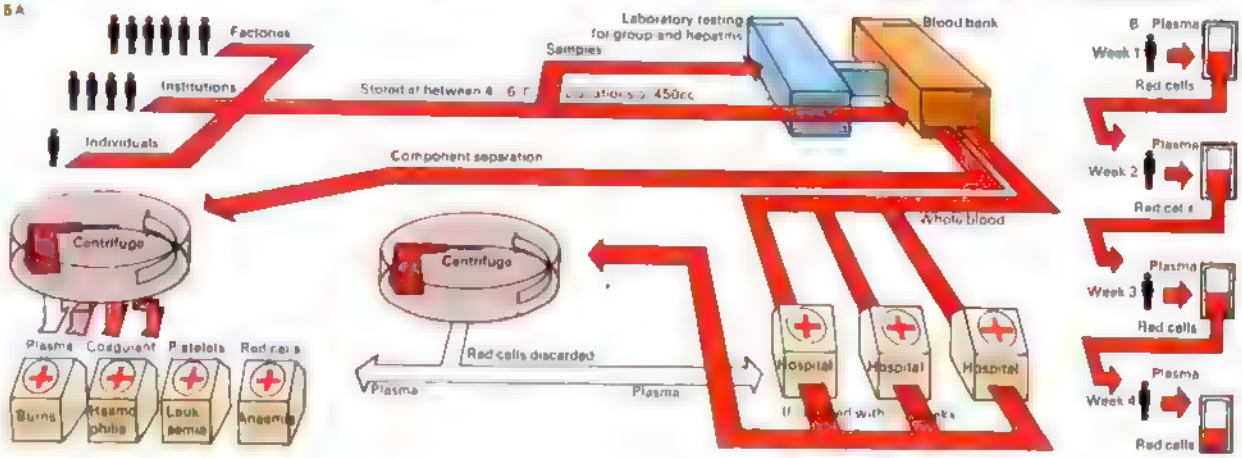


2 The first "purpose-built" operating tables [A] appeared in the 19th century. They were cumbersome but basically the same as modern ones. The ether inhaler [B] replaced whisky or rum as a pain killer.



3 The discovery by Pasteur that germs cause infection led to Joseph Lister's introduction of antiseptics – the treatment of wounds with carbolic acid (phenol). Lister (1827–1912) first introduced a hand-operated spray, then a steam mode. The steam [1] from the boiler [2] – heated by methylated spirit on a wick [3] – sucked carbolic acid vapour from a jar [4], thus filling the air with an antiseptic mist.

4 Wars accelerate all technologies, not least that of medicine, and the carnage of World War I caused great strides in surgery, particularly in bringing damaged tissues together and keeping patients alive with a saline solution, in spite of heavy fluid losses. Here, in a makeshift field operating theatre, an army surgeon removes a bullet from the arm of a wounded soldier.



5 A constant supply of fresh blood from donors [A] to a blood bank is critical for it deteriorates after three weeks. Blood reaches hospitals whole (at 4°–6°C [39°–43°F]) or split into its constituent parts by centrifuge. Once separated and dehydrated, plasma can be frozen and kept indefinitely as powder. To increase plasma supplies [B] a donor can give blood 40 times a year instead of the normal three, receiving back his red cells from the previous visit.

dressing to the stitched incision dramatically reduced the death-rate after surgery [3]

Revolutionary and beneficial as it was, antiseptic surgery began to give way, at the start of the twentieth century, to the modern practice of aseptic surgery, in which the aim is not to kill germs in the operating theatre but to exclude them as far as possible from the patient. Surgeons developed an aseptic (germ-free) technique in which the operation site is made bacteria-free by swabbing the skin with a bactericide before an incision is made. Sterile towels, gowns and masks were introduced and the surgeon and his assistants learned to scrub their hands carefully for a full five minutes before every operation. As an additional precaution sterile rubber gloves, first developed in the USA by William Halstead (1852–1922), became customary.

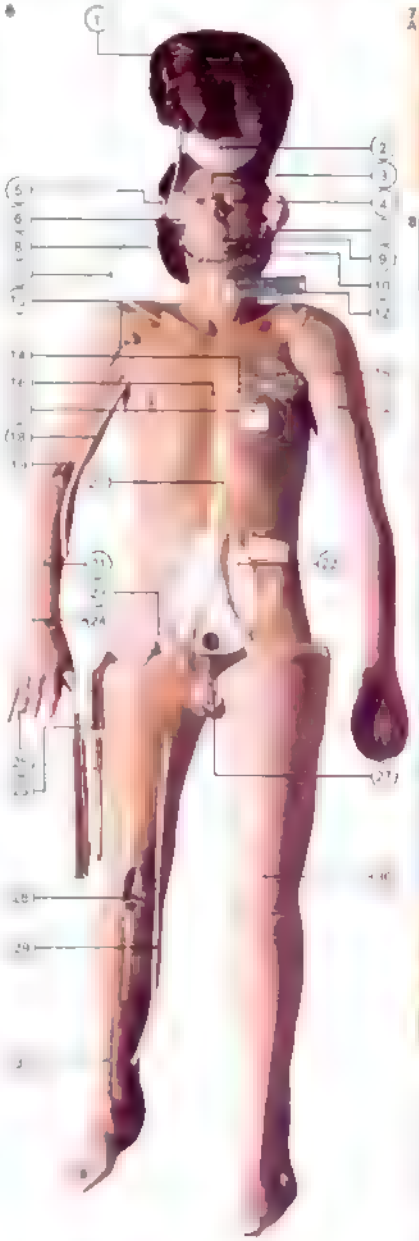
The 1920s saw the widespread introduction of blood transfusion [5] which made it possible to replace blood lost during an operation and, at the same time, there were refinements in anaesthesia to enable opening of the chest for lung operations.

Shortly after World War II chest surgeons found ways of inserting a finger into the heart in order to clear a partially obstructed valve. Next it became possible to lower an anaesthetized patient's temperature (hypothermia) so much that his heart could actually be stopped for a short time, thus permitting even more delicate operations within it.

Modern innovations in surgery

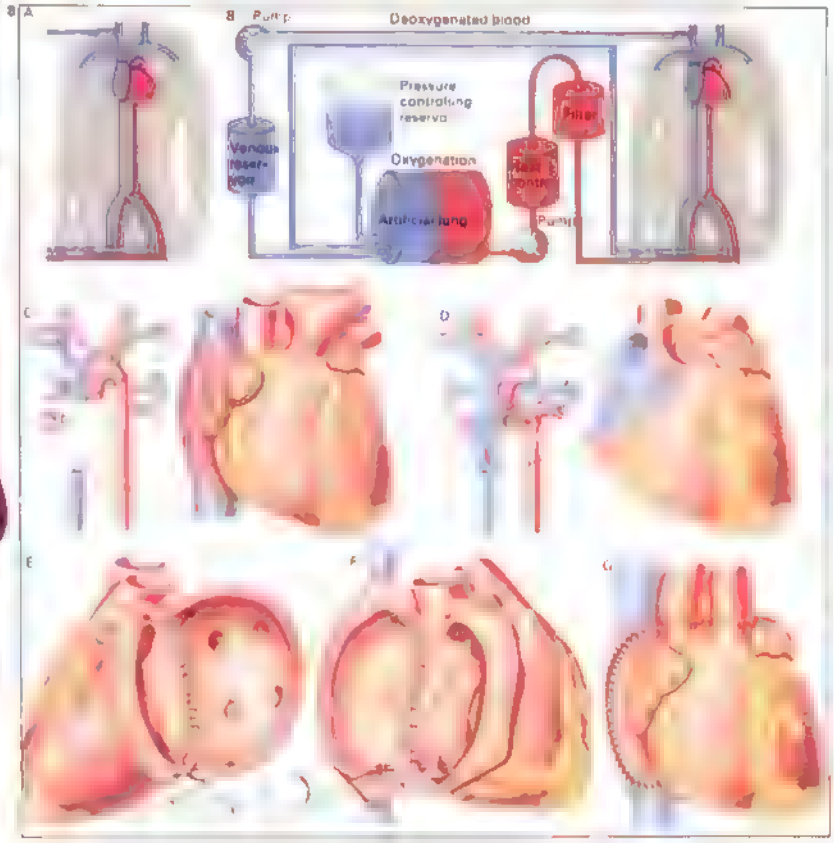
The advent of the heart-lung machine (which completely takes over the functions of pumping the blood and of breathing) meant that surgeons could take far more time over their complex procedures (with correspondingly better results) and eventually progress to the transplantation of hearts [8]. By the 1960s and 1970s it became possible to transplant other organs (kidneys, liver, lungs and pancreas), although the incidence of failure is still very high. Indeed, surgery and its technology have advanced to such an extent that it is now possible to conceive of a situation in which anyone who has a damaged, defective or worn out organ will be able to get a working replacement, real or artificial, as well as purely cosmetic operations [6].

Surgical instruments of the 18th and 19th centuries looked crude but were effective. Basic amputation saws (A, B), knives (C, D) and scalpels (E, F, G) have hardly changed. That's mechanical saw (H), in which the handle was rotated to move the blade was sometimes used. Today, clamps have replaced tourniquets (I), while forceps (J) now resemble scissors more. The trocar (K) was a sharp, hollow tube used for drawing off fluids or tissue.



6 "Spare parts surgery" has undergone dramatic changes in recent times, making it possible to transplant living organs and fit artificial substitutes within the body. New metals, acrylic and other plastics have helped overcome the problems of rejection, the refusal of the body to accept the installation of foreign materials. Possible spare parts include [1] surgical wig [2] "Vitallium" skull plate, [3] plastic nose implant, [4] cosmetic acrylic eye, [5] in-the-ear hearing aid, [6] silicon plastic ear, [7] cosmetic plastic ear, [8] metal jaw bone, [9] dentures, [10] chin implant, [11] Spitz-Holter valve to control fluid on the brain, [12] electronic larynx, [13] shoulder joint replacement, [14] heart valve replacement, [15] heart pacemaker, [16] filter preventing blood clotting in circulation to the lungs, [17] Dacron heart patch, [18] Dacron artery replacement, [19] elbow replacement, [20] Dacron vein and artery graft, [21] metal bone plate, [22] plastic replacement after removal of a part of small intestine, [23] hip joint replacement, [24] wrist bone replacement, [25] finger joint replacement, [26] thighbone replacement, [27] cosmetic plastic testicles, [28] knee joint replacement, [29] plastic artery graft, [30] artificial leg with knee and ankle movement, [31] shin, [32] artificial arm

7 In basic plastic surgery, the burned area (A) is cleaned and "postage stamps" of skin from elsewhere on the body are grafted on (B). Healthy skin grows out until the whole area is covered (C).



8 A heart transplant involves the removal of a badly diseased heart, replacing it with one from a "donor" — someone who has just died, usually as the result of a road accident. At first, heart-lung machines are connected to both the donor (A) and the recipient (B), to keep their hearts alive. The donor heart is then removed (C) and the recipient's heart is cut out (D) leaving only the top-most part. The donor's heart is now stitched on to this "stump", first on one side (E) and then on the other (F). Salt solution is then pumped through the heart to clear it of air and the final stitching to the great vessels is completed (G). The recipient's new heart is now stimulated into beating by means of an electric shock. The operation is extremely rare.

Having an operation

Although advances in modern techniques have made such well-publicized operations as organ transplants and brain operations relatively safe, these are still only a minute fraction of the total number of operations performed each year. There are, for example, only about 500 kidney transplants performed each year in England and Wales out of more than 260,000 surgical operations.

Common surgical operations

To a certain extent operations are a matter of fashion. In the United States circumcision was, until recently, the most common operation. Virtually the entire male population underwent circumcision shortly after birth. In other countries, however, it is becoming quite rare because it is only really necessary in relatively few cases.

In the United Kingdom and many other countries appendectomy (the removal of the appendix) is probably the operation most frequently performed by general surgeons. Other common ones include herniorrhaphy (the repairing of a hernia, or rupture), partial gastrectomy (the removal of part of the stomach, usually for ulcers), cholecystectomy (gall bladder removal), vagotomy (cutting of the nerves to the stomach, again for ulcers); mastectomy (breast removal), and haemorrhoidectomy (removal of piles).

The general surgeon may spend a great deal of his time removing benign and malignant tumours from various parts of the body and, in many hospitals, be expected to strip out varicose veins. The latter is also in the province of the vascular surgeon, who carries out operations on blood vessels, such as the removal of aneurysms (swellings) in arteries.

Orthopaedic surgeons are responsible for operations on bones, joints, tendons, ligaments, nerves and muscles. The common procedures they undertake include meniscectomy (removal of a damaged cartilage from the knee joint) and operations for the relief of "slipped discs" in the vertebral column.

Urological surgeons deal with the urinary tract and the male reproductive organs. One of their main operations is prostatectomy or removal of the prostate gland. Gynaecological surgeons deal with the female reproductive organs and carry out the extremely common operations of hysterectomy (removal of the womb) and "D and C" (dilation and curettage, or widening of the neck of the womb and scraping of the lining).

Other surgeons specialize in operations on the ear, nose and throat; the most common operation they perform is probably still tonsillectomy, or excision of the tonsils. Considerably less frequent are the delicate operations carried out by eye surgeons, brain surgeons and plastic surgeons.

The procedure for surgery

Most people have some sort of operation (though often only a minor one) in the course of their lives, but few understand very much about what is going to happen to them [1]. On admission to hospital the patient is carefully examined by a house surgeon (a junior doctor working for the specialist) to ensure that he is fit for an operation. The house surgeon may order blood tests or X-rays and, if the operation is to be a major one, may ask the blood bank to put aside some blood that has been carefully "cross-matched" with the patient's blood.

CONNECTIONS

See also
Development of surgery
Health, nursing and

1 Before an operation a pre-medication injection [1] is drawn up into a syringe and administered [2], by a nurse. The "pre-med" takes effect after half an hour or so and then a hospital porter transfers the patient to a trolley [3] and wheels him into the anaesthetic room [4]. The anaesthetist [5] "induces" the patient with an anaesthetic drug [6]. Administering it into an arm vein makes the effect swift, although sometimes the patient is put to sleep by inhaling gases from the anaesthetic machine. This is a movable trolley carrying cylinders [7] of oxygen – which has to be administered throughout the operation to keep the patient alive – and anaesthetic gases such as cyclopropane. Oxygen and whatever gas has been chosen by the anaesthetist are mixed together and they flow through a black rubber bag [8]. When the patient can no longer breathe for himself, this bag is squeezed manually by the anaesthetist or his assistant. After an initial intravenous injection the anaesthetist will continue anaesthesia with gas as the injection wears off. In the anaesthetic room, a "drip" [9] may be set up if it is felt that the patient may need fluid or blood replacement. A blood-pressure

At some time during the 24 hours before the operation the patient is asked to bathe thoroughly and the skin around the operation site is then shaved. For about six hours before, no food or drink is allowed since, at the start of the anaesthetic, such food might be vomited and inhaled into the lungs. An hour or so before the operation a nurse gives the patient a "pre-med" injection, which makes him feel calm and relaxed and dries up chest secretions that might interfere with his breathing under the anaesthetic.

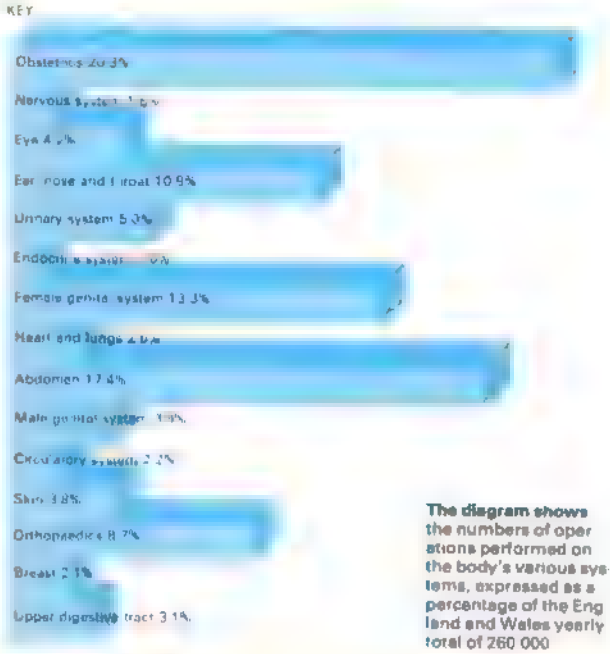
Eventually the patient is taken to the operating theatre suite. Here the anaesthetist gives him an injection into a vein, this produces complete unconsciousness in about ten seconds. The patient is now wheeled into the operating theatre and placed on the table. The anaesthetist places a mask over his face and keeps him unconscious with a mixture of anaesthetic gases and oxygen, often supplemented by injections.

The surgeon, his assistants and the theatre sister take up their positions. They clean the operation site carefully and place sterile towels round it. The surgeon then makes his incision and probes down to the organ or tissue he is looking for, while his assistants use small forceps to close off any bleeding points. The theatre sister's function is to pass instruments and swabs to the surgeon. She is backed up by a team of more junior nurses and by theatre technicians.

Most operations are over in about an hour. At the conclusion, the surgeon carefully stitches up the wound and applies a dressing. The patient is then taken to the recovery room where the anaesthetist (or a nurse) looks after him until he awakens. From there he may go to the intensive care ward, but more usually he will be returned to the ordinary ward, where he will be observed carefully for a few hours.

The recovery period

The patient will spend a few days in hospital after surgery if the operation has been a minor one. Stitches are usually removed about a week after the operation. After discharge from hospital the patient usually sees the surgeon at least once more before being left in the care of his own family doctor.



The diagram shows the numbers of operations performed on the body's various systems, expressed as a percentage of the England and Wales yearly total of 260 000

cuff (10) and monitoring devices may be applied to his arm before he is finally wheeled (11) into the operating theatre. Meanwhile the surgeon, assistant surgeons, theatre sister and other theatre staff have been scrubbing up (12) and dressing

With the patient on the table (13) they commence the operation while the anaesthetist (14) monitors the patient's condition. The theatre sister (15) passes instruments (16) to the surgeon while other technicians and nurses (17) adjust lights, check blood pressure and brain rhythms (18) and put swabs on a scale (19) to check blood loss. After the operation the patient may need to stay in an intensive care room (20) where machines (21) monitor his vital functions. He should

soon be able to move on to a recovery room or a small observation ward (22) which is equipped with oxygen, suction equipment and drip-feed (23). Here the nurse (24) will keep an eye on him until his pulse and breathing are satisfactory

Dentistry

The 5,000-year-old history of dentistry reflects not only man's desire to preserve his teeth but also his attempts to avoid and relieve the agonies of toothache. Despite advances made over the centuries, and the highly scientific nature of dentistry today, most people still have dental problems – mainly teeth decay and gum disease. The statistics are alarming, showing, for instance, that in 1968, 36 per cent of the population of England and Wales over the age of 16 had none of their natural teeth.

Dental decay

The teeth are covered by a gelatinous bacteria and food-laden layer [2C] called dental plaque. Decay or caries begins with the breakdown of carbohydrate foodstuffs caught in the plaque by bacteria normally present in the mouth. The resulting strong acids dissolve away a protective layer of enamel overlaying the dentine of the teeth [1]. Once the enamel is penetrated, destruction of tooth tissue is rapid.

It is known that decay can be eradicated almost completely by the use of simple and effective preventive measures. These include fluoridation of water supplies, control of dental plaque and advice on diet.

The action of fluoride on teeth, in strengthening the enamel and increasing resistance to decay, was first recognized less than 40 years ago, although it has benefited people for centuries in places where water supply naturally contains fluoride. Since the 1930s, addition of fluoride to water supplies in selected areas at a level of one to two parts per million has significantly reduced the occurrence of dental decay.

The effectiveness and ease of administration of this form of fluoridation make it attractive, although other measures can be used. They include "painting" the teeth with strong fluoride solutions, adding fluoride to toothpastes and mouthwashes, and "fissure sealing" (a technique whereby susceptible parts of the tooth are sealed by the application of a plastic coating).

Preventive efforts must obviously be directed at children, for teeth, unlike other tissues in the body, cannot repair themselves once they have been damaged. When decay sets in, older established methods of dental treatment in the form of fillings, crowns, extractions, bridges and dentures must be used. The repair work done by a dentist aims to remove all decay and to stop further destruction by reconstructing the tooth using a filling material – either metal or plastic [4].

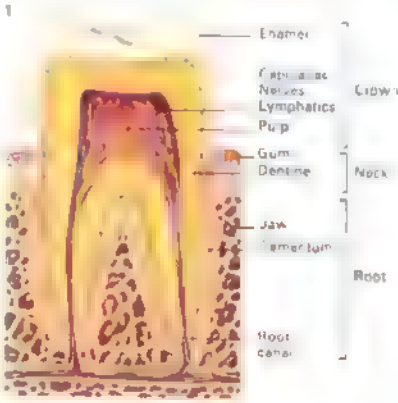
Where dental decay is so far advanced that little of the tooth remains, crowns may be required. This involves removing nearly all of the tooth above gum level and building up in gold or porcelain [5]. Porcelain crowns may also be used to improve the appearance of misplaced or discoloured front teeth.

More serious damage

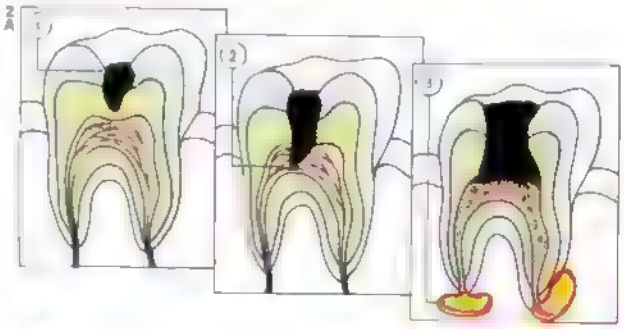
While decay can continue for a time quite painlessly, it will eventually cause toothache, which becomes acute once the decay penetrates the nervous and vascular tissue (the pulp) in the centre of the tooth. By this stage an ordinary filling may not be enough to save the tooth, for unless the bacterial inflammation of the pulp can be eradicated, an abscess will develop around the root of the tooth and cause its loss. Teeth with inflamed pulps or

CONNECTIONS

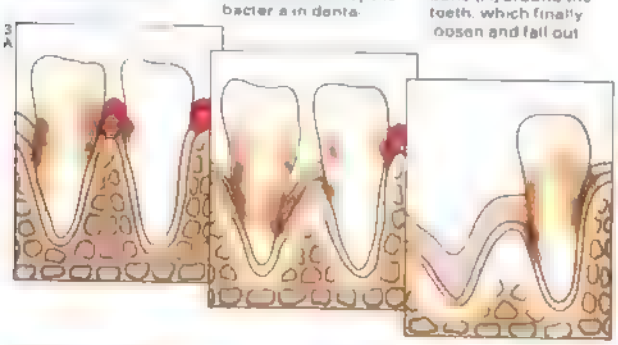
See also



1 The three main parts of a tooth are the crown, the neck and the root. The crown, capped by enamel, projects out of the gum. The root is held in the bone of the jaw by a ligament embedded in cementum. The delicate pulp contains nerves and blood vessels which connect with the bloodstream and main nerve paths via the root canal. The root and crown underlying the enamel are formed from porous dentine.



2 Tooth decay (caries) occurs because sugar and other carbohydrates caught in dental plaque (a yellowish film that sticks to the teeth) [C] are turned into strong acid by the fermenting action of bacteria in the mouth. Under the layer of plaque the acid destroys the tooth first dissolving away the enamel and then eating into the softer dentine, leaving dark, disfiguring stains [B]. Decay [A] starts by attacking the crown [1]. If untreated, it progresses down until it affects the pulp [2]. Once in the pulp, the bacteria begin spreading down the root canal and an abscess develops in the bone around the root, causing pain and swelling [3].



3 Gum disorders are dealt with by periodontology. Gum recession [B] follows infection and inflammation caused by the bacteria in dental plaque. If untreated this leads to loss of bone [A] around the teeth, which finally loosen and fall out.

4 To repair decayed teeth, a dentist may use several different types of filling [A]. In the most common procedure, the affected part is cut away [1] and the cavity sealed, usually with a metal filling such as silver amalgam in a back tooth [2] or a white plastic filling in a front tooth. When decay has spread very close to or into the pulp a special sedative lining is placed under the filling to protect the pulp and prevent toothache [3]. If the pulp is badly decayed, a root treatment may be used, all the damaged pulp is removed and the root canal is sealed with a special cement [4]. The effectiveness of good dentistry is shown by a case before [B], and after [C] treatment. Decay and a broken down incisor were repaired using silicate fillings and a crown on the incisor.

abscesses can sometimes be saved by a technique known as root treatment in which the entire contents of the root canal are removed and replaced by filling material

As a last resort, a tooth can be extracted under a local anaesthetic given by injection. (Where there is an abscess a general anaesthetic is sometimes used.) The tooth is firmly gripped by a pair of specially designed forceps and moved in such a way that the surrounding bone yields gently and the tooth can be pulled out without damage. When wisdom teeth remain buried, or a tooth is badly broken and decayed, extractions involve cutting back the gums and removing some of the jaw-bone surrounding the tooth Such surgical extractions are usually carried out under general anaesthetic

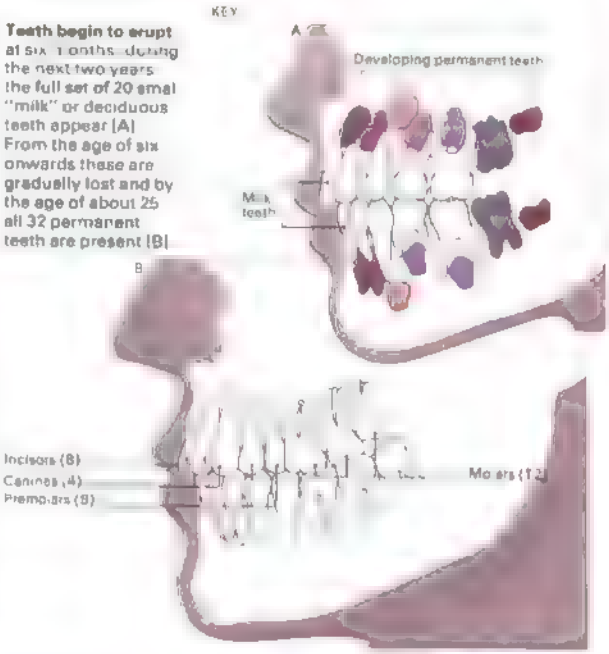
When a tooth has been lost it is often desirable that it is replaced, not only for appearance and to make chewing easier, but also because once a gap exists other teeth may have impaired function Such gaps can be filled with dentures [6] or bridges [7].

As with dental decay, periodontal or gum disease can usually be prevented by conscientious oral hygiene However, even when gum disease is established, with loosening of the teeth and gum recession, effective treatment is available [3] The gums are cut back, the damaged tissue removed and the patient told how to prevent recurrence by the regular and thorough practice of oral hygiene

Specialized forms of dental surgery

One of the more specialized branches of dentistry is orthodontics, which includes the realignment of tooth positions using spring-loaded appliances (braces) within the mouth Such treatment is most effective in childhood [8]. Orthodontics also extends into the field of general surgery, being used to treat cleft palates that prevent patients from eating and talking normally [9] Repairing fractures to the jaws and facial bones, correcting facial deformities, and treating tumours within the mouth also require special dental skills

The scope of dentistry is extremely wide While one of modern dentistry's concerns is the prevention of caries and gum disease, much can be done to help patients once disease or injury exists



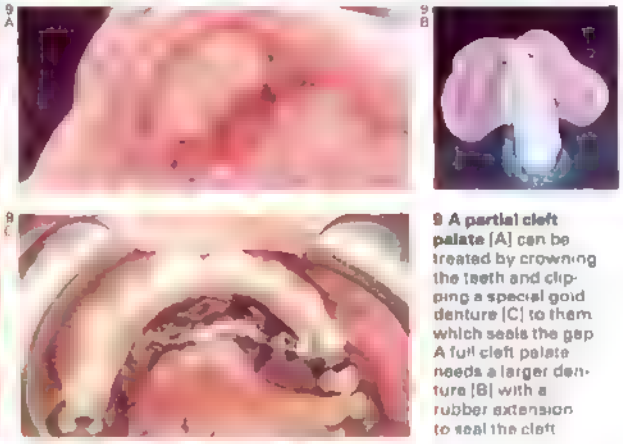
Teeth begin to erupt at six months during the next two years the full set of 20 small "milk" or deciduous teeth appear [A] From the age of six onwards these are gradually lost and by the age of about 25 all 32 permanent teeth are present [B]

5 Crowns – gold for back teeth and porcelain for front teeth – are used to repair badly damaged teeth Where decay has made an ordinary filling impossible [A], the tooth is cut down to a peg and the dentist takes an impression so that an accurately fitting crown can be made This is then cemented on to the remaining tooth stump Even where a tooth has broken off at gum level [B], it can still be restored by first fitting a gold post into the root of the tooth [C] and then using this as a support for the crown [D] Crowns are also used for cosmetic purposes to replace unsightly discoloured or misshapen front teeth

6 A denture may be the only solution if too many teeth in one jaw are lost Dentures are fitted not only for cosmetic reasons and to allow chewing, but also to maintain sufficient teeth to prevent rapid deterioration of those that remain and imbalance in the muscle action of the jaws, which can cause spasm and pain The making of an accurately fitting denture is a skilled job A wax impression is taken of the mouth and a model made to serve as a mould for casting the denture The extent of a denture varies as do the materials that can be used to make it Plastic [A] and metals such as gold or chrome cobalt [B] are common

7 A missing tooth can be replaced by a bridge construction [A] The teeth on either side of the gap are cut down to a peg shape so that gold crowns can be fitted over them. A unit of three crowns, one of which replaces the missing tooth, is then cemented into position. If necessary, a bridge can also be made using two teeth on the same side of the gap [B] One of the teeth used to hold the bridge was badly decayed [C] and needed root treatment and the fitting of a gold post [D] before the bridging crowns could be made and fitted [E]

8 Orthodontic braces are used to realign misplaced teeth such as protruding upper incisors. A plate clips into the mouth and stainless steel springs are used to move the teeth gently into the right line.



9 A partial cleft palate [A] can be treated by crowning the teeth and clipping a special gold denture [C] to them which seals the gap A full cleft palate needs a larger denture [B] with a rubber extension to seal the cleft

Birth control

All human societies have faced the problem of balancing the need for more hands with the difficulty of feeding more mouths. And most have tried to regulate fertility for personal or economic reasons. Before scientific knowledge of birth control a wide range of methods was used either to prevent conception or to limit intercourse by ritual taboos.

Development of birth control

In simple societies early methods were often based on sheer superstition. To ward off unwanted pregnancies women wore "magic" amulets made from the tooth of a child, for example, or the testicle of a weasel [3].

Other early methods of controlling conception reflected some knowledge of human biology, or at least close observation of cause and effect. Medical recipes for counteracting sperm date back to about 2000 BC. Vaginal douches of varying efficacy were used, as were simple versions of condoms to sheathe the penis or diaphragms to block off the uterus. The "safe period" of women was recognized by a number of primitive peoples and for hundreds of years men practised the oldest method of all, coitus interruptus (removal of the penis before ejaculation).

Hostility towards "artificial" methods of birth control led to heated controversy during the nineteenth century when urban overcrowding and soaring birth-rates brought the issue of birth control to the attention of social reformers [Key]. The courageous work of women such as Aletta Jacobs (1849–1929), Margaret Sanger (1883–1966) and Marie Stopes (1880–1958) brought gradual acceptance of family planning by women.

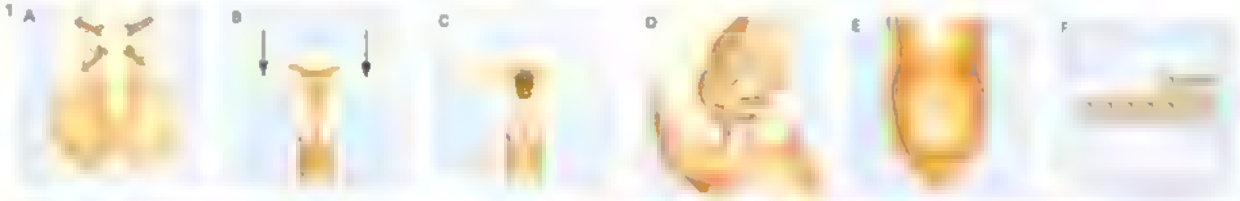
The advent of the Pill

It was not until the middle of the twentieth century that the emphasis switched from attempts to stop fertilization to methods of intervening in the production of fertile eggs and sperm. In 1955, an American team of biologists led by Gregory Pincus (1903–67) discovered that the hormones oestrogen and progesterone, when taken orally, were highly effective in preventing ovulation [2]. By 1966 ten million women throughout the world were estimated to be taking the Pill, the number today may be as much as 40 million.

The success rate of the Pill is such that women have been able to enjoy intercourse without the fear of unwanted pregnancy and without having to employ barrier methods. Described by some authorities as a safer drug than aspirin, the Pill is effective and generally easy to use. It does, however, have some side effects and its use is not advisable for women with a history of blood-clotting, diabetes, cystic fibrosis, liver disease or cancer of the breast or reproductive organs. In addition since it decreases normal vaginal secretions, venereal diseases are more readily transmitted. Although evidence about its long-term effects is still limited, fears about loss of fertility seem unfounded. One study has shown that 60 per cent of women formerly on the Pill became pregnant in their first cycle while only 40 per cent of other women conceived as quickly. Some women on the Pill also report increased sexual desire, although reactions are mixed. There are about a dozen "Pills" with varying amounts of oestrogen and progesterone and finding the most suitable one for a particular woman may require

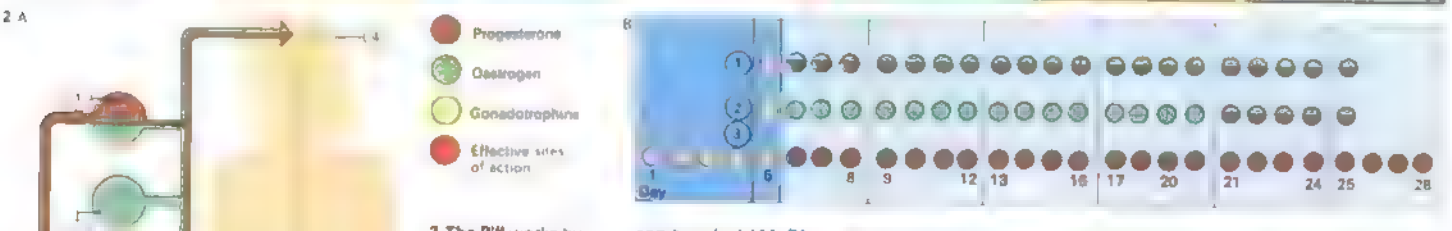
CONNECTIONS

See also



1 The main methods of birth control are vasectomy [A], the Pill [B], the intrauterine device [C], the cap [D], the condom [E] and the rhythm method [F]. Coitus interruptus (withdrawal of penis before ejaculation) is practised but not reliable and often causes undue strain, both physical and emotional. Spermicidal creams, foaming tablets and aerosol sprays are equally unsatisfactory on their own, as is the douche – an attempt to wash out sperm from the vagina. The belief, once widely held, that a woman could not conceive while breast feeding is not true. Prostaglandins, hormone-like substances now under study, can theoretically be used to induce menstruation in a pregnant woman but such a method of "birth control" could be seen as a form of abortion. Female sterilization (cutting and tying the Fallopian tubes) is effective but often irreversible.

How it works Ducts taking sperm from testicles to penis are cut and tied. There is no effect on ability to enjoy intercourse.	Taken every day for most or all of the menstrual cycle, the Pill inhibits ovulation and/or hinders penetration of the sperm into the cervix.	No one yet knows exactly how the IUD works. But it is thought either to inhibit implantation or to affect the sperm in some way.	All types of cap, inserted into the vagina and placed over the cervix, provide a barrier to sperm, preventing them reaching the womb.	The condom or sheath, rolled on to the penis just before intercourse, provides a barrier to sperm, preventing them from reaching the vagina.	Basically, a couple refrain from intercourse when the woman is most likely to conceive (at least 10 days each month and probably longer).
Possible effects There need be no psychological problems. Physically, the operation is not risky but there is no guarantee it can be reversed.	Migraine, weight gain, depression, mid-cycle breakthrough bleeding, nausea, plus possible extra risk of thrombosis. Change of Pill often helps.	Heavy periods do sometimes occur and removal of the intrauterine device may be required. No long-term harmful effects are known.	Generally, caps are thought to be harmless, except in rare cases where there is an allergy to rubber or plastic or sensitivity to a chemical used.	Condoms are harmless – unless the user or his partner is sensitive to rubber. Some couples also find condoms lessen sensation.	Some couples find that restricting intercourse in this way imposes severe strain and tension on the relationship.
How to use A doctor or clinic will advise on surgeons who carry out this operation. It does not usually involve a stay in hospital.	The right variety must be prescribed by a doctor or family planning clinic. Periodic check-ups advisable on grounds of general health.	The IUD must be fitted by a gynaecologist. Regular check-ups will be advised. Many IUDs have threads enabling their position to be checked.	Caps have to be fitted either by a doctor or the family planning clinic. To be effective they must be used together with a spermicidal cream.	Perhaps the most widely sold contraceptive available over the counter. It should only be put on when erection occurs.	Careful records of menstrual cycle and temperature needed. Difficult to use if periods are irregular or when sickness affects temperature.
Success rate Tests are carried out after a few months to check seminal fluid is sperm-free. The chance of conception is about one in 30 000.	The risk of becoming pregnant is almost nil provided the Pill is taken exactly as directed. Failure rate 1 in 300 woman-years.	IUDs have been known to come out unnoticed with the result that in a few cases pregnancies have occurred. Failure rate is 1 in 20 woman-years.	Caps are thought to be about as effective as the condom and are dependent on the care with which they are used. Failure rate 1 in 8 woman-years.	Needs care in use. Is safest with a spermicide. Some condoms have been known to leak. Failure is about 1 in 8 woman-years.	Timing cannot be guaranteed, so the calendar method is risky, but temperature is more reliable. Failure rate 1 in 4 woman-years.



2 The Pill works by means of the contraceptive effect [A] of hormones similar to natural oestrogen and progesterone. Three main types are widely available using synthetic forms of these hormones with varying dosage schemes. [B] The combined pill [A1, B1] has both oestrogen and progesterone and one pill is taken each day for about three weeks in each monthly cycle. The hormones act together on the anterior pituitary gland [4] to inhibit normal production of gonadotrophic hormones and thus suppress ovulation [5]. Even if an egg is released the Pill acts further to prevent pregnancy by affecting the oviducts [6] and womb lining [7] and by altering cervical mucus [8]. During the fourth week withdrawal of hormones produces a lighter and often shorter menstrual period beginning within three or four days. A month's supply of sequential pills [A2, B2] consists of oestrogen taken alone for 15 days and then a progesterone oestrogen combination for the following five days. The mini-pill [A3, B3] in contrast, relies on direct progesterone effect only. Dosage is continual but there is a normal ovarian cycle and menstruation takes place as usual. (the first five pills are placebos) although the cycle may be more variable than normal.

a considerable amount of experimentation

Various other methods of birth control remain popular and in most cases reasonably effective – these include vasectomy intrauterine devices (IUDs), the cap, the condom and the rhythm method [1] Choice of a birth control method is influenced by availability, religious attitudes and personal aesthetics as well as safety and effectiveness The most suitable method is not only a matter of personal choice but also of medical advice

Outlook for developing countries

Researchers are investigating the possibility of a pill that could be taken only once a fortnight and of long-lasting hormone injections or implants that could prevent ovulation (the last would be particularly useful in developing countries) As well as the possibility of a post-coital pill, scientists may soon develop a pill, to be taken by men, that would suppress the production of sperm Psychologists are doubtful however whether women would readily trust men to take a pill as directed

Family planning has been most effective in the developed countries where a strong force for the widespread use of contraception has been the Women's Liberation Movement, which stresses the woman's health welfare and status In the Third World, where the threat of starvation is so much greater the struggle to persuade people to limit family size has unfortunately been less successful [4] One problem is that the message of population control comes from former colonial powers and is sometimes seen as an attempt to retain dominance In addition, children are seen variously as symbols of virility (especially in South America) a justification of a woman's life, an insurance policy against old age and a work force

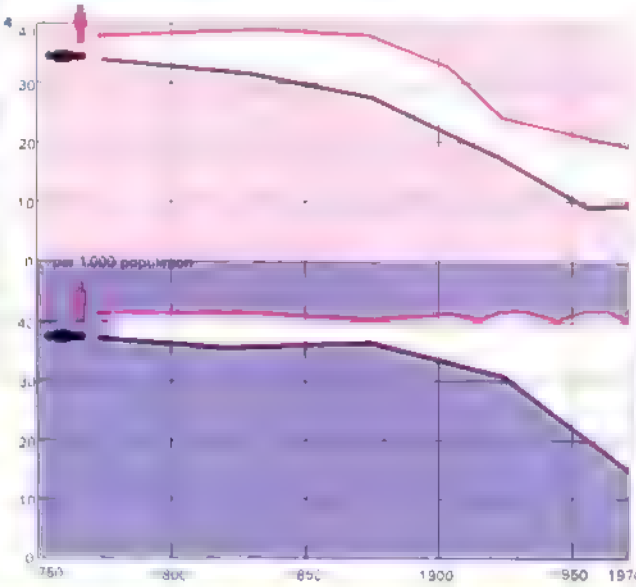
The problem of birth control exists not only on the biological front but also as a serious social dilemma in a world of limited space and resources. Education alone seems to offer a long-term solution [5] The Indian government, which offers incentives for sterilization [6], considered a law demanding compulsory vasectomy for every male with three or more children The ensuing outcry indicates the improbability of such a law

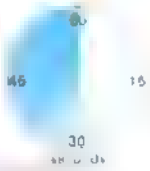
Large families like this were common in the West at the end of the last century but are rare now Family planning began in The Netherlands with the establishment in 1882 of a clinic founded by Aletta Jacobs Use of a diaphragm – the "Dutch cap" – was explained at the clinic but this kind of instruction in birth control made slow headway elsewhere Margaret Sanger set up the first clinic in America in 1916 but it was closed soon afterwards and Mrs Sanger was arrested Similar public hostility including charges of obscenity was faced by Marie Stopes who opened the first British clinic in 1921

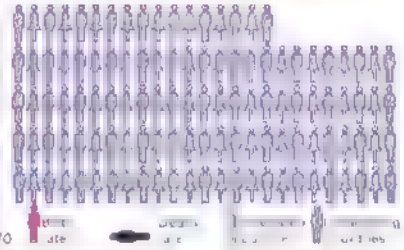
3 Early methods of contraception included the wearing of lucky charms [A] the drinking of supposedly magic potions [B] and even a resort to fervent prayer [C] The chastity belt [D] sought to prevent intercourse Casanova used a lemon [E] as a spermicide while the Egyptians used the [illegible] of a crocodile [F] Early condoms [G] were made of sheep or pig intestines

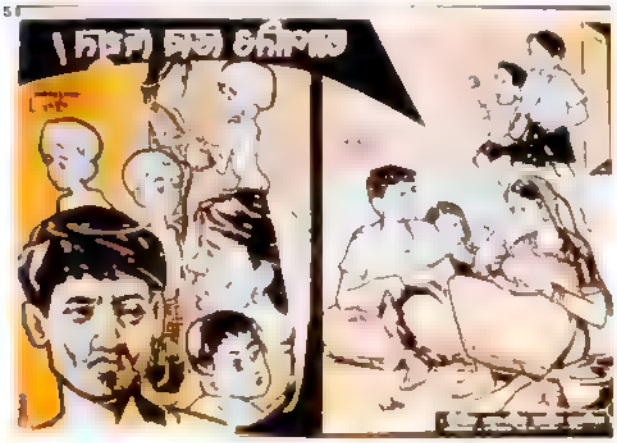
4 World population reached 1 000 million by 1850 But by 1970, due largely to a sharp decline in mortality rates it had overshot the 3 500 million mark Statistics show that in developing countries in Africa Asia and Latin America and parts of Oceania the birth rate is far higher than in more industrialized countries of Europe North America Japan Russia temperate South America New Zealand and Australia This high birth rate combined with a fall in deaths due to better nutrition and medical facilities has meant that the annual rate of increase in population is now far greater in those countries whose ability to provide food or adequate social services is already most strained The aim of population control is to reduce this strain on resources and to promote better health and happiness

The graphs show birth and death rates per 1,000 population in developed and developing countries since 1750 The silhouettes beside them represent the current growth in population every 60 seconds









5 Family planning posters are used in India to publicize the advantages of smaller and healthier family units But some Indians said "Look at the unlucky man with only three children"

6 After sterilization, a group of men from Kerala in southern India were given an array of gifts – in airline bag a plastic bucket, an umbrella a lottery ticket and cash – all incentives in an intensive local campaign for birth control

Looking in on the unborn foetus

A neurone, the basic working unit of the human nervous system. It took a magnification of 600 times actual size to produce the photograph from which this image was reproduced.

Introduction to mental health

Psychiatry, derived from two Greek words meaning "mind" and "medical treatment", is that branch of medicine devoted to the diagnosis, treatment and prevention of mental ill health. There is no clear dividing line between physical medicine and psychiatry. Psychiatry cannot safely ignore the relationship between bodily condition and mental state, and medicine is becoming increasingly attentive to the manner in which a person's emotional state can precipitate and aggravate physical ill health.

The causes of mental illness

There is never a simple or single cause for mental illness [2]. In practice, a distinction is drawn between predominantly intrinsic or endogenous causes, such as inherited factors, and predominantly extrinsic or exogenous causes, such as physical injury and disease and mental stress such as bereavement, unexpected financial reverses or the loss of a job. The interaction between a person's basic personality and any physical or emotional stress affects the extent of any subsequent mental reaction. One man or woman may be able to cope with massive stress quite adequately, whereas another may be overwhelmed by what seems a minor set-back.

Disorders of the mind

The symptoms and signs of mental illness are grouped under a number of headings: disorders of perception; of thought and speech, of memory, of emotion, of the experience of the self, of consciousness, and motor disorders.

Hallucinations are sensory perceptions without any external, objective stimulus and are a common perceptual disturbance in psychotic illnesses and brain disorders. Symptoms of thought disorder include flight of ideas, in which the thought processes are speeded up; perseveration, whereby the patient "perseveres" with a particular response long after a change in his immediate environment has demanded a different response, and thought blocking, in which there is a sudden arrest of the train of thought and the start of an entirely new one.

In certain disorders (obsessional states for example), the patient recognizes that he is compelled to think about certain things despite his every effort to rid his mind of them. Such compulsion is understood by the patient to originate within him and not as a consequence of some alien activity. In thought alienation, however, the patient experiences his thoughts as being under the control of some external agency. He may believe that others are participating in his thinking or that thoughts are being inserted into or extracted from his head, or that others think his thoughts in unison with him and are aware of his most intimate contemplations.

A delusion is defined as a false belief of morbid origin, and which is an absolute conviction unamenable to reason or contradiction although usually absurd or even impossible. Disorders in the form of thinking are characterized by a fragmentation of the links connecting successive thoughts, and the phenomenon of over inclusiveness, in which the patient is unable to maintain the boundaries of a concept.

Disorders of memory include disturbances in the registration of information (such as by lack of concentration) and in the retention and recall of material. Where there is a

CONNECTIONS

See also

1 Compared to a "normal" person's view of a scene [A] the neurotically anxious person may feel threatened by his immediate surroundings [B]. Nevertheless, he will maintain contact with reality and be aware that his surroundings are in fact unaltered. The psychotic, however, may experience the same situation in a seriously distorted way [C], believing his impaired perception of external reality to be a valid and objective one.

2 Mental health is influenced by widely different factors. Genetic make-up is important, identical twins [A] have a much higher chance of both suffering from schizophrenia than normal siblings. Genetic factors are less significant in neuroses in which emotional trauma, particularly in childhood [B] is more important. Sexual difficulties are traditionally linked with psychiatric disorders, particularly in adolescence [C]. Social isolation [D], to which immigrants and the elderly are vulnerable, causes depression. Some jobs have attendant psychiatric risks [E]: alcoholism in journalists and seamen, depression in housewives and semi-skilled workers, and psychosomatic complaints in doctors and businessmen. Some physical disorders like hypothyroidism [F] are associated with psychiatric symptoms and it is difficult to separate them from a purely mental condition such as anxiety neurosis.

pathological memory loss, the gap may be filled with elaborate fabrications, as in alcoholic psychosis

Emotional disorders consist of variations in the intensity or duration of the emotional response, which may also be inappropriate to the particular situation. The term "affective disorder" refers to a sustained disorder of mood, such as depression or mania, rather than to a transient emotional reaction

Depersonalization, a disturbance in the experience of the self, occurs when the individual feels himself changed in comparison with his former state. He feels like an automaton and watches his own actions "from outside". In derealization the person perceives the outer world as strange and altered in some significant way. Both these disturbances are experienced on occasions by perfectly healthy people

Disorders of consciousness are mainly due to physical causes and include alterations in attention and concentration, a slowing in thinking and a lack of direction in thought and action. The patient may be disoriented, may manifest disconnected and incomprehensible behaviour or may be delirious

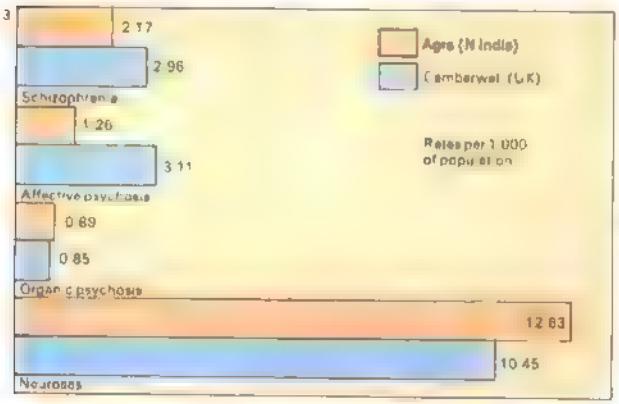
Motor disorders include lack of initiative, retardation in speech and action and stupor. Major forms of motor disorder seen in mental illness are psycho-motor activity, seen often in manic states, catatonic excitement, characterized by stereotypy (monotonous repetition in speech, mannerisms and movement); and passivity feelings (whereby patients believe their impulses or feelings are controlled by some outside agency)

Major categories of mental illness

There are three major categories of mental illness. The psychoses [1] consist of schizophrenia, manic-depression, paranoid illnesses, the organic psychoses and psychoses associated with physical disorders, the neuroses, consisting of anxiety and phobic states, obsessive-compulsive disorders, hysterical and depressive neuroses; and the personality disorders, which include alcoholism and drug dependence, the behavioural disorders of childhood and a group of anomalies or deviations of personality that are not the result of a psychosis or any other illness

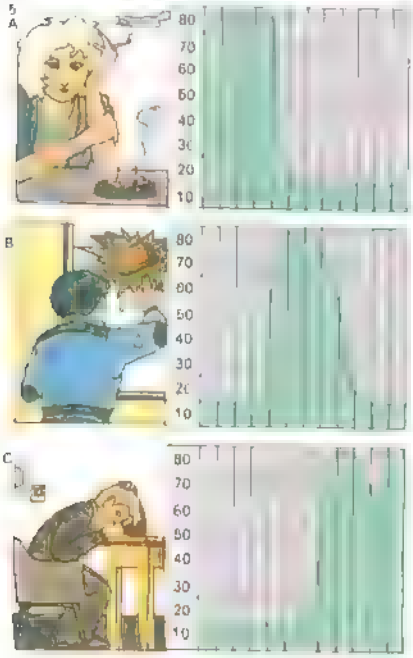
KEY

The blank, emotion less expression of a shell shocked soldier shows a typical response to overwhelming stress



3 Variations in the prevalence rates of psychiatric illnesses between different countries may reflect genuine difference. However they may be the result of differing diagnostic practices among psychiatrists or of a marked variation in the availability of psychiatric care. For this reason it is difficult to assess whether one country or one environment is more liable to encourage a particular category of mental illness than another. There do, however, appear to be variations in types of neuroses. For instance, the commonest neurosis in Agra is hysteria whereas in Camberwell it is depression

4 Men shows cyclic variation in many of his functions, as do almost all plants and animals. The most prominent cycles are the circadian or diurnal rhythms [A], which are about 24 hours in length and vary little over a long period (inset clocks represent six weeks). The best known are the diurnal fluctuations in body temperature, steroid secretion and sleep. These rhythms can be disturbed in various ways. In constant darkness they shift forwards so that sleep [B] occurs later and later. Extreme disturbance of sleep can lead to psychotic conditions. Jet travel across international or national time zones [C], where the body is having to readjust its clocks to local time, can produce a number of psychiatric symptoms. The outer clock represents local time, the inner clock the patient's sleep time



5 Responses to a set of questions may give indications of the presence and degree of psychiatric illness. The questionnaire is designed to elicit different responses from people with different mental illnesses or disorders of behaviour. Questions relate to various areas of behaviour or diagnostic categories, such as hysteria, phobia or thought disorder. Answers are grouped into such categories and the score on each category registered on the graph. The result is a characteristic profile for different mental states. Those shown here are for a neurotic patient [A], a boy suffering from a behaviour disorder [B], and a psychiatric patient [C]

How mental illness has been treated

In ancient times an individual with a disturbed mind was assumed to be influenced by spirits, demons, gods or other supernatural forces. As a result healing practices were generally unsystematic and relied heavily on the power of suggestion

Madness through history

The theory of the four humours, proposed by the Greek philosopher Hippocrates (*c* 460–377 BC) and Galen (*c*. AD 130–*c*. 200), introduced a degree of order and influenced medical thought for over 2,000 years. This theory postulated the existence of four key elements in the body (blood, phlegm, yellow bile and black bile) Different diseases and different personality temperaments were believed to be associated with the predominance of one or other humour An excess of black bile or dry phlegm, for example, was considered the cause of melancholia and this condition was treated with physical methods, such as vapours, baths, diet and emetics

Arab medicine continued the traditions inherited from Greek medicine and between the eighth and thirteenth centuries a number of asylums for the insane were opened at Damascus, Cairo and Baghdad. In Europe, with the spread of Christianity, care of the mentally afflicted was one of the duties of monasteries, other religious houses and hospitals run by the clergy. The first hospital to be built exclusively for the insane is believed to have been opened at Valencia, in Spain, in 1409

Paradoxically, this greater care and compassion for the mad among the religious institutions coincided with a tendency among ordinary people to reject them along with paupers, cripples and other social outcasts Added to this there was medieval Europe's obsession with sorcery, heresy [3], witchcraft and demonic possession. Many violent or dramatic outbursts of insanity were attributed to these "evil forces" and dealt with by torture, imprisonment and often death Yet the most popular treatise of the time, *De proprietatibus rerum*, written by an English Franciscan friar, distinguished between physical and psychological causes of mental illness and prescribed rest, sedation and music therapy for the violently disturbed

The seventeenth and eighteenth centuries witnessed a remarkable growth in the scientific basis of medicine and surgery Theories about mental disturbances, however, lagged behind and the fashion to explain madness in terms of moral flaws, lack of impulse control and the degeneration of personality flourished. Attempts were made, nonetheless, to explain how the mind worked in physiological and chemical terms, even if these did take unusual forms

Theory and therapy

Franz Mesmer (1734–1815) blamed mental illness on the accumulation of a magnetic fluid in the body that could be removed by special magnetic powers possessed by certain therapists, such as himself

Around this time, Franz Joseph Gall (1758–1828) claimed to have discovered 27 organs within the brain, each of which was responsible for a particular mental function The better these organs worked, the larger they were and the more they affected the overall shape of the skull. Phrenology, the art of judging character and mental stability by

CONNECTIONS

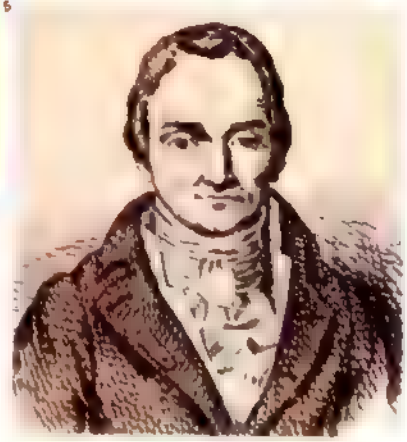
See also

1 Evidence of trephination, the removal of a part of the skull probably as a treatment for mental illness, has been found in many cultures dating back at least 10,000 years It was believed that the spirit causing the mental distress might be released through the opening in the skull Later the medieval theories of demoniacal possession provided another theoretical basis for such a therapeutic approach to be readopted

2 Pilgrimages to the shrines of special saints were made by the mentally ill during the late 12th century One of the most famous shrines was that of St John at Ghent A complex ceremony of offerings, processions, penances and prayers lasted nine days This led to the founding, in 1181, of the first reception centre for vagrants, paupers and the insane, which was paid for by the citizens of the town

3 Insanity has often been linked with magical practices, sorcery and religious heresy Girolamo Savonarola (1452–98) the Florentine religious and political reformer seen here being burnt to death was almost certainly mentally unhinged There were few medical authorities, however who distinguished between insanity and heresy The court physician of the Duke of Cleves, Johann Weyer (1515–88) was prominent among those who insisted that almost all the so called witches were elderly women who were psychotic rather than evil or possessed but neither the Church nor legal experts agreed It was not until the latter half of the 17th century that the general persecution of witches ceased

4 Benjamin Rush (1745–1813) is generally regarded as the father of American psychiatry A controversial figure, he was a vigorous advocate of blood letting and the use of mechanical devices in the treatment of mental illness He is better remembered for his progressive views on the need to combine psychological and physiological approaches in the study and treatment of individual patients



5 Philippe Pinel (1745–1826) was the French physician who on 24 May 1798 with the consent of the Legislative Assembly, removed the chains from 49 insane patients at the Bicêtre Hospital in Paris, ushering in a new type of treatment

6 William Tuke (1732–1822), an English wholesaler of tea and coffee and a philanthropist, founded with his son Henry (1755–1814) the York Retreat (opened 1796) This became an influential early institution for the humane care of the insane

the shape and features of the skull, became widely popular until it, too, fell into disrepute around the mid-nineteenth century.

Meanwhile the era of confining the insane had dawned and large mental hospitals and private madhouses multiplied. Conditions in these institutions were often appalling and treatments relied on physical methods [7].

Towards modern psychiatry

In time, a number of progressive pioneers removed the chains, and methods of treatment based on kindness, understanding, limited freedom and work began to supplant those elaborate mechanical restraints previously in vogue. At the same time two contrasting trends, both present in amorphous forms for several centuries before, began to crystallize. The first, initiated mainly by German diagnosticians, led to the gradual introduction of an orderly system of classifying mental disease. The second involved the discovery of the unconscious.

The French psychologist Pierre Janet (1859–1947) described a hierarchy of mental functions at the bottom of which he placed automatic functioning, such as in catatonic stupors, and at the top, rational, experienced and conscious activity. Janet's theories, together with the elaborate demonstrations of hypnosis by Jean Charcot (1825–93) at the Salpêtrière in Paris and the remnants of the theories of mesmerism, magnetism and phrenology provided the background to the theories of Sigmund Freud (1856–1939) concerning the unconscious.

The twentieth century has witnessed the development and popularization of psychoanalytical theory, particularly in the United States. At the same time the discovery of the therapeutic efficacy of electrical treatment in severe depressive illnesses, the development of powerful anti-psychotic drugs, such as chlorpromazine, and the slow unravelling of genetic aspects of some forms of mental illness illustrate the progress made in comprehending the biological contribution in psychiatry. The necessity for understanding and effective countermeasures is illustrated by the fact that in Britain one man in every nine and one woman in every six can expect to be treated for mental disorders.

KEY

Bedlam, the popular name for the Bethlem Royal Hospital, became infamous for its ill treatment of the insane. Public torture was common and a penny tour through the crowded cells was a fashionable pastime in 18th century London.

7

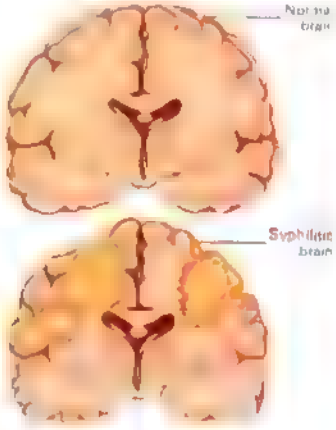
A Plunge bath
B Revolving bed
C Whirling cage
D Revolving cage
E Tranquilizer chair

7 Elaborate mechanical contrivances superseded, in the 19th century, the physical methods used in earlier periods. The purges, emetics and bleedings. With the simple intention of subduing violent and manic patients, the most extraordinary contraptions were devised. Leather straps, canvas jackets, muffs and handcuffs replaced the chains. Patients were fastened in wooden chairs in such a way that only the minimum of movement could occur. The induction of sudden and intense fear was believed to have a beneficial effect in mental illness with the result that physicians employed whirling chairs, spinning beds and padded hollow wheels into which patients were strapped and which rotated at speeds of up to 100 revolutions per minute. Such rotation produced vomiting, incontinence, loss of consciousness and, in a number of cases, death. Another cruel device was the plunge bath. The unsuspecting patient was invited to cross a floor which was opened beneath him plunging him into a bath of ice-cold water. Benjamin Rush, in America, tried to call attention to the need for therapy and the employment of suitable companions to listen sympathetically to patients.

8

8 General paralysis of the insane is a dreaded complication of syphilis appearing some 10 to 25 years after the primary infection. The condition is characterized by severe personality changes, social uncontrollability, impulsiveness, grandiosity and serious mental deterioration. During the latter half of the last century, before the organic cause had been discovered, physicians looked for psychological causes. The Japanese microbiologist Hideyo Noguchi (1876–1928) [A] put paid to imaginative speculations when, in 1911, he discovered at post mortem the syphilitic spirochete in the shrunken atrophied brains [B] of patients dying from general paralysis of the insane. His discovery led to a resurgence of hope that similar distinct and treatable causes might be found for other mysterious psychiatric disorders.

B



Psychoses

A psychosis is a behavioural disorder typified by a fundamental break with reality and a tortured internal experience [Key]. The major psychosis is schizophrenia. This term refers to the severe mental disturbances, common to all forms of the disorder, in which there is a detachment from the external world and a disintegration of the internal one.

Schizophrenia

Schizophrenia usually, although not invariably, first occurs at puberty. The most common symptoms include incoherent thinking, lack of emotional response, delusions and hallucinations [1]. Thought processes are jumbled and the resultant speech is often uncomprehensible and illogical. The schizophrenic may receive moving news without any sign of emotional response. He may smile or appear indifferent when told of some personal tragedy or when confronted by a sombre scene. Delusions are often bizarre and occur with a strength of conviction that renders them beyond reasonable argument. A schizophrenic in a bar, on being offered a drink by the barman, might conclude that this was a signal to the other customers that he was a murderer. Hallucinations, particularly involving voices, are also common in schizophrenia.

The cause of the condition is obscure. The role of genetics, in the light of recent studies of twins, is undoubtedly important but it is clear that, although necessary, genetic factors may not be sufficient for the illness to appear. Certain emotional stresses such as family disharmony, physical precipitants such as infection or childbirth, and a number of drugs including amphetamines and LSD, may trigger the illness in predisposed people.

Schizophrenia may be sub-classified as paranoid (the patient has delusions of persecution and is mistrustful of everyone), simple (he is apathetic, withdrawn and uncommunicative), hebephrenic (the patient exhibits silliness and bizarre behaviour); and catatonic (he may sit for hours in one position with muscles rigid, unable to feed or look after himself, seemingly unaware of anything around him). Schizophrenia commonly shows as a mixture of these types. Treatment includes drugs (phenothiazines), social rehabilitation and psychotherapy. The illness usually has serious consequences, and although some patients do recover, permanent damage is the more common outcome.

Paranoia and manic-depression

Paranoia is used by some psychiatrists to refer to mental disorder that occurs when someone has a permanent and unshakeable system of delusions and, at the same time, complete clarity and order in thought, action and will in all other areas. The condition is rare – many cases that begin as paranoia end up showing the characteristic widespread intellectual and emotional deterioration of schizophrenia. However, paranoid illnesses do occur in some solitary and shy individuals who project their own doubts and insecurities on to others, in a number of deaf people, who may misinterpret comments they do not hear clearly as insulting, and in the morbidly jealous.

The other serious psychosis is called manic-depression. As implied, there are two main forms, mania, characterized by excitement, euphoria, grandiose schemes, rapid

CONNECTIONS

1 Schizophrenia is characterized by disjointed and episodic thought [A] and inappropriate emotional responses [B]. There is a tendency to delusion with the reading of innocent gestures as malicious signals [C]. Hallucinations and illusions occur as when a peaceful country scene seems oppressive or terrifying [D].

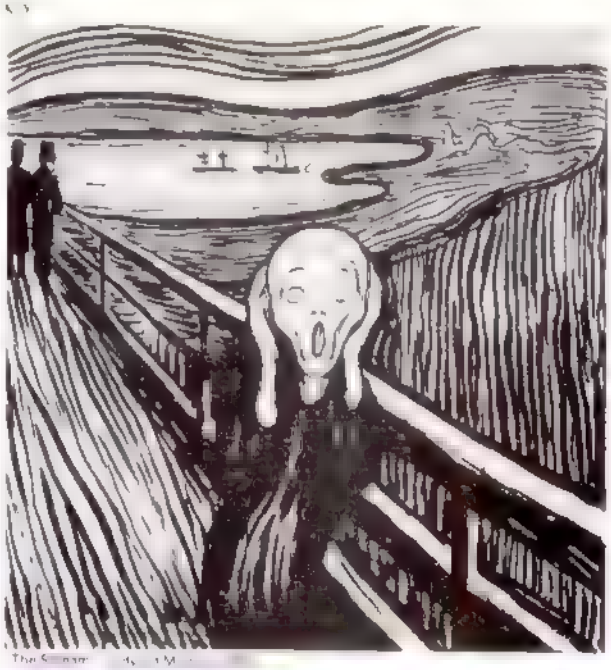
flights of ideas and overactivity; and depression [2]. Patients may suffer from one or other type or from both.

Psychotic illnesses often occur in association with organic diseases. Many different poisons and various kinds of brain damage may affect the mental state. Psychotic symptoms commonly occur in delirium associated with fever, drug intoxications and other causes of brain disturbance. Patients in severe delirious states have illusions and hallucinations of the senses, especially vision. Dislocations of their sense of time and space also help to differentiate the organic psychoses from the so-called functional psychoses schizophrenia and manic-depression.

Brain diseases and psychosis

The most serious brain disorder often appearing with psychotic features is dementia. This is a diffuse cerebral disease in which there is serious intellectual deterioration and emotional impairment. The brain shows general atrophy, there is a loss of brain cells and specific changes in the cerebral cortex. The psychosis may take various forms – manic, depressive or paranoid. The manic variety is characterized by pointless activity, silly boasting and a marked overtalkativeness. In the depressive variety, the emotions are rather blunt, there is obvious irritability and hysterical symptoms may be intermingled with hypochondriacal ones. The paranoid variety is most dramatic: there are often elaborate delusions of gases being pumped into rooms, food being poisoned, spies trying to kill the sufferer, and so on.

Dementia occurring before the age of 50 is commonly termed presenile. Pick's disease, in which there is a circumscribed atrophy of the frontal lobes, and Alzheimer's disease, in which there is widespread atrophy of the brain, may be accompanied by psychotic features. Psychosis can also occur in cerebrovascular disease with brain tumours and as a consequence of head injuries. Epilepsy, particularly affecting the temporal lobe, may be accompanied by hallucinations, mood disturbances and paranoid delusions. In the period immediately following childbirth psychoses can occur in predisposed women.



2

2 Depression, in psychiatric terms, refers to a condition in which depression is not the only symptom but is accompanied by insomnia, loss of weight and appetite, diminished libido, feelings of worthlessness and guilt and suicidal thoughts. Some depressive states may have an obvious precipitating factor such as a bereavement or a broken engagement ("reactive" depression). In others, and particularly in patients predisposed genetically to develop manic-depressive psychosis, the mood disturbance often occurs with little or no obvious provocation ("endogenous" depression). In extreme depression the patient may be severely retarded or markedly agitated. He tends to be preoccupied with feelings of guilt and may explain these by concluding that he has committed some terrible crime or sin for which he is being punished. Depressed patients commonly feel a burden to others, particularly to those close to them, and may attempt suicide to relieve not only their own feelings but also those of their concerned relatives and friends from whom they invariably feel cut off and alienated. A delusion sometimes seen in severely depressed people is the so-called "nihilistic" delusion. In fluenced by the profound emotional malaise that is central to depression, the patient's mood may become completely negative. He may declare that he has no name, no age, no parents, relatives, wife or children. He may insist he has no head, no chest, no body indeed, he may deny everything and resist everything. Treatment in the severe forms of depression is by electric shock and antidepressants. Psychotherapy is helpful particularly in reactive depression.

Personality defects and neuroses

The neurotic person has obvious abnormal psychological symptoms, but does not show that sharp break with reality that characterizes the psychotic person. Particular neuroses can and do appear for a short time in otherwise healthy people, but they are usually exaggerated versions of normal experiences. All of us have felt depressed, anxious, fearful at some time – what distinguishes the neurotic is not a qualitative difference from the norm but a quantitative one

Types of neuroses

Classification of neuroses into various categories is fairly arbitrary but is based on the major symptom present. The main neurotic states identified are anxiety neurosis, hysteria, depressive neuroses and obsessive - compulsive disorders

The overriding complaint in anxiety neurosis is an excessive fear, often amounting to panic and associated with physical symptoms such as a dry mouth, palpitations and sweating. It may be complicated by a phobia [3, 7], obsession [4] or depression [6]

In the popular mind, hysteria is linked with uncontrollable tantrums and screaming fits. For the psychiatrist the term refers to those symptoms, both physical and mental, such as paralysis, tremor and amnesia, that are caused by a psychological disturbance and aim, unconsciously, at escape from a seemingly insurmountable difficulty or the fulfilment of some need

Hysteria takes two main forms known as conversion or mainly physical disturbances [5] and dissociation The most common of the dissociative states is the fugue when someone quite suddenly and without any warning signs wanders off. Such a journey has no plan or destination, is usually accompanied by loss of memory, and often allows the person to avoid some unpleasant situation, such as an embezzlement charge or divorce

Allied to fugues are the so-called trances and twilight states when a person withdraws and insulates himself from the world More rare but of the same type is the dual or multiple personality, when a person manifests a number of traits that are opposed to each other and poorly integrated By a process of dissociation a shy, prudish girl may become flirtatious and seductive. In extreme cases people keep different names and styles for their different personae and when playing one role deny all knowledge of any other.

Anxiety, neurotic depression and hysterical neurosis are the reactions to stress of more or less neurotic personalities. The strongest and best-balanced individual may react with depression to a severe set-back or with anxiety to a major stress. However, most anxious, depressed or hysterical patients are more the victims of their own personalities than of outside events

Personality disorders

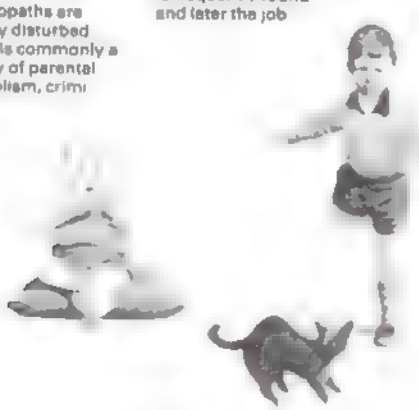
Personality disorders are a group of anomalies or deviations which, although not the result of a psychosis or other illness, are odd enough to upset or puzzle others and sometimes even the sufferers themselves. Such disorders resemble mental illness and may require psychiatric help and understanding but the sufferer is less ill than abnormally developed, in the same way that a physically handicapped individual is not ill but has an abnormal physical constitution

CONNECTIONS

See also

1 A psychopath displays recurrent anti social delinquent and criminal behaviour in many areas of his life Data on the prevalence of the disorder are unreliable but it is seen frequently by psychiatrists, usually because of an associated alcohol problem or depression The family backgrounds of most psychopaths are grossly disturbed There is commonly a history of parental alcoholism, criminality, separation, divorce or early death The first signs usually occur in late childhood or early adolescence and are commonly a restlessness, an unresponsiveness to discipline and a tendency to be cruel to smaller children or animals. A disturbed school history, with truancy and academic failure, is frequently found and later the job record is marred by poor performance, unreliability, constant changes and an inability to accept even minor criticism. Psychopaths have great difficulty in maintaining close emotional relationships and their marriages are characterized by infidelity, separation and divorce Most psychopaths end up with criminal records Many of them indulge in minor petty delinquency but a small minority commit brutal and callous acts of physical violence

1



2

2 Inadequate individuals show a low physical and mental strength a lack of resilience and flexibility and an inability to cope with stresses They are often nervous and dependent in early childhood and their parents are either intensely protective or harsh and rejecting They are excessively shy, socially and sexually inhibited, egocentric and introverted They often lead lonely, somewhat joyless and anxious lives in adulthood

Some personality disorders are due to faulty brain development caused by genetic make-up, injury, infection, poisoning or malnutrition in childhood. Some sex chromosome anomalies also seem to be associated with antisocial behaviour, in particular the so-called XXY male who is unusually tall, aggressive and sometimes of subnormal intelligence. However, environmental factors, including early childhood deprivation, parental quarrelling and severe stress, are also known to play their part.

A number of specific categories of personality disorder, including the antisocial [1] and the immature [2], have been described. The paranoic is sensitive and vulnerable, reacting to everyday experiences with an excessive sense of inferiority and humiliation, and apt to be touchy in preserving what he conceives to be his rights. People who have so-called affective personalities have long-standing anomalies of mood – either a predominantly gloomy and pessimistic attitude to life or the opposite. In contrast the hysterical personality has shallow and changeable emotions and, although unreliable himself, craves love and attention. The schizoid is notably aloof, shy and reserved, tending to be markedly introspective and eccentric. Finally, the anankastic or obsessive personality is characterized by a strong sense of insecurity, an excessive caution combined with a stubborn inflexibility and a rigid perfectionism.

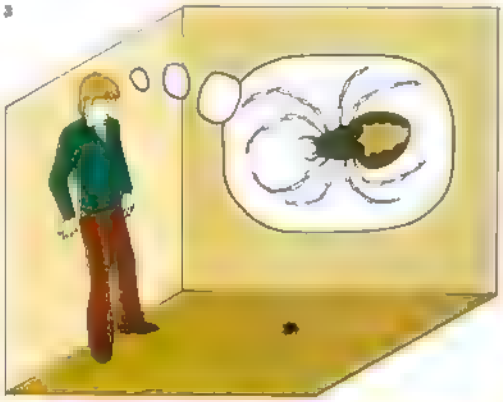
Sexual problems

Various forms of sexual problems are referred to psychiatrists. Some, such as fetishism (the exclusive derivation of sexual pleasure from inanimate objects) are private anomalies of sexual behaviour. There are others, however, such as paedophilia (the desire of an adult to engage in sexual activity with children) and exhibitionism (the need to expose the male genitals to females) that involve public conduct.

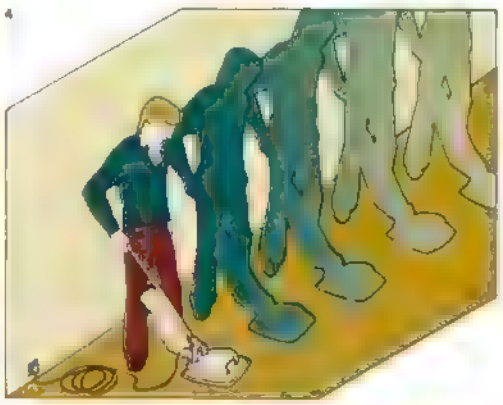
Transvestism (dressing in the clothing of the opposite sex) and trans-sexualism (the wish to be and to function as a member of the opposite sex) are rare but important. These problems are regarded as developmental anomalies rather than as true illnesses.

KEY

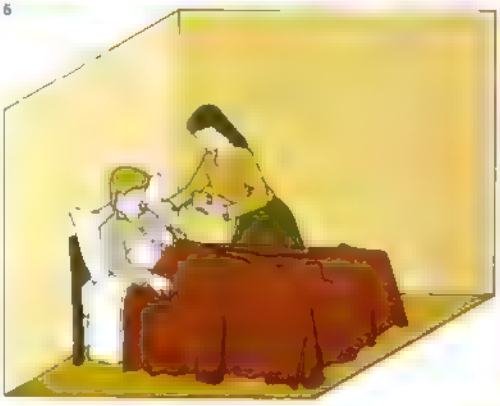
The Munchausen syndrome, named after Baron von Munchausen (1720-97), is a severe personality disorder wherein a person malingers or consciously feigns an illness to obtain a desired end. A few people repeatedly malinger, going from one hospital to another often under a variety of assumed names, but almost always telling the same story, faking the same symptoms and submitting to innumerable investigations and operations. Chronic malingerers resemble the Baron, who entertained his friends with extraordinary (and patently untrue) tales of his supposed travels so that he eventually gained the reputation of being an incorrigible liar.



3 Anxiety neurosis may be complicated by a phobia or dread of a specific object or situation. A common phobia is that of 'creepy-crawlies' especially spiders. Mildly phobic individuals experience anxiety only when required to handle spiders but the severely phobic can become terrified just by thinking about them and will avoid any situation in which there is the slightest chance of further exposure.



4 In obsessional neurosis the sufferer is occupied with thoughts that do not really interest him and impulses that seem alien. He feels impelled to perform actions which not only give him no pleasure but which he is powerless to stop. The compulsive rituals may be eccentric (eg touching objects a given number of times) or involve repetitive activities (eg cleaning) aimed at neutralizing morbid thoughts (of dirt).



5 A common feature of hysterical neurosis is conversion, a physical disturbance psychologically caused. Symptoms are paralysis, tremor, fits, blindness, deafness and sensory disturbances. Conversion is a primitive mechanism for dealing with difficult situations and evoking sympathy. Hysterics differ from malingerers in that although they may produce illness they are unaware of what they are doing.

6 Depression is a reaction to the stress or frustration produced by some incident like a loss, grief or disappointment. Depressive neuroses, as a rule, are an exaggerated form of what happens when the average man is temporarily cast down. The severe loss of weight, appetite and sleep, the agitation, delusions and deep guilt feelings seen in manic depressive psychoses are nearly always lacking.

7 Agoraphobia, a fear of open spaces, is commoner in women and includes panic experienced when out alone, shopping or travelling. Other phobic stimuli are heights, fire, lightning, death and hair. Fears of flowers, water and numbers are rarer, but fears of social situations are common, while children are often afraid of the dark. Snakes, closed places (claustrophobia), cats, spiders, insanity, dirt and mice are other well known causes of phobia.

Psychosomatic diseases and retardation

Certain physical disorders such as hypertension and bronchial asthma [3] have classically been designated psychosomatic since the early 1930s. In the intervening years, many other conditions have been similarly labelled including anorexia nervosa, obesity, psychogenic vomiting, some abdominal pains, diarrhoea, torticollis, diabetes mellitus and writer's cramp.

There is little agreement on which illnesses are or are not psychosomatic. Their only common characteristic is the assumption that psychological factors play a major role in their complex and obscure causes. Some critics believe the term to be misleading in that the description of some illnesses as "psychosomatic" implies that the remainder are not. In practice probably few illnesses do not include psychological factors.

Psychosomatic diseases are believed to be the result of a chronic and exaggerated state of the normal physiological expression of emotion, affecting a particular vulnerable organ. If persistent, such an emotional reaction can often produce structural damage.

There has been much speculation about the possible link between certain personality types and susceptibility to psychosomatic disorders. It has been claimed, for instance, that some sufferers from peptic ulcers may be people who unconsciously want to remain dependent. Such a wish is productive of low self-esteem, however, and runs counter to the adult ego's pride and desire for independence. Accordingly, it is repressed and compensated for by aggression and ambition. The resulting inner conflict produces chronic anxiety which in turn leads to overproduction of gastric juices and so to ulcers.

In anorexia nervosa [1] an unwillingness on the part of the patient to develop womanly characteristics, both physical and psychological, with their attendant social and sexual expectations, is thought to lie at the heart of the refusal to eat, the fear of obesity and the desperate attempts to maintain an abnormally low weight level.

Mental retardation

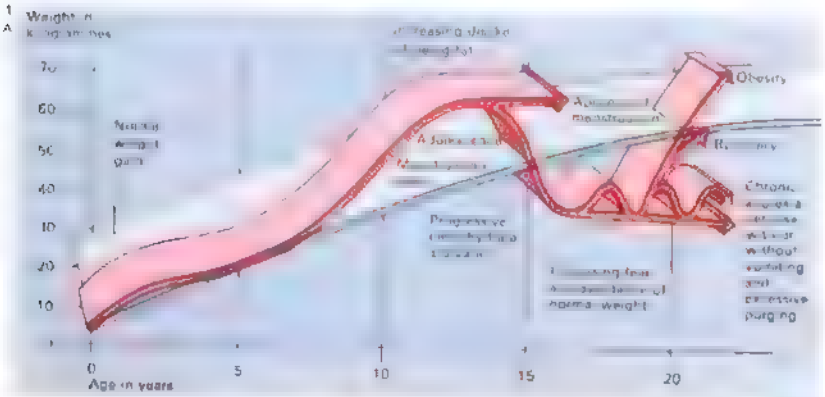
Mental retardation refers to intelligence defects from birth or before full brain development irrespective of the cause. If the IQ (intelligence quotient) of a large sample of people is assessed [7], the same kind of normal distribution curve results as when physical characteristics, such as height or weight, are studied in population samples. IQs between 90 and 110 are considered average. Mental retardation is divided into a number of grades: borderline (IQ 68–85), mild (IQ 52–67), moderate (IQ 36–51), severe (IQ 20–35) and profound mental retardation (IQ under 20).

Such a grading acts as a rough guide to the capabilities of people falling within each range in terms of possible social adjustment and functioning, ability to learn and the acquisition of skills. For example, an adult with profound mental retardation requires nursing care and may achieve only a very limited level of self-care. The moderately retarded adult may be able to work in an unskilled or semi-skilled capacity under sheltered conditions, whereas the person who is only mildly retarded can achieve adequate social and vocational skill for a minimal level of self-support.

Many factors at different stages can cause

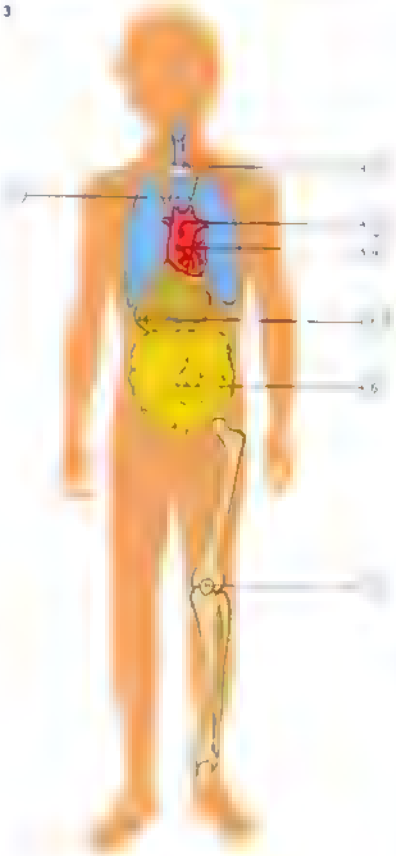
CONNECTIONS

See also

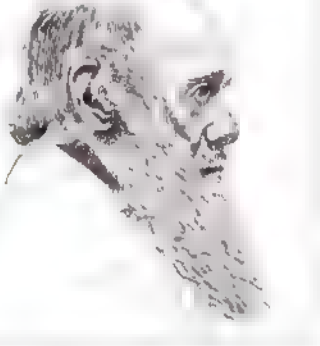


1 Anorexia nervosa a rare condition occurring mainly in young women, sets in about the start of puberty (A). It begins with distaste for, or a phobia about, fatness that soon develops into an inability to eat, often exacerbated by the taking of emetics. The result (B) is low body weight, absence of monthly "periods", constipation and downy hair over the normally hairless parts of the body.

2 The "vapours" were a fashionable psychosomatic complaint of distinguished ladies of the 18th century, characterized by dramatic fainting and a variety of nervous fits, sometimes attributed to over tight corsetting. Famous physicians of the day treated attacks of the vapours with such "modern" treatments as electric shocks or a course of baths. Impressionable young women were frequent victims.



3 Psychosomatic diseases are believed to be the result of excessive and prolonged exposure of a vulnerable organ to the normal physiological changes that are caused by emotion. If persistent, such an emotional reaction can produce structural damage in the organ affected. Major physical diseases in which psychological factors are thought to play a significant causal role, in addition to modifying the way in which such illnesses develop, include bronchial asthma (1), thyrotoxicosis (resulting from overactivity of the thyroid gland) (2), hypertension (3), coronary artery disease (4), peptic ulceration (5), ulcerative colitis (6) and rheumatoid arthritis (7). There has also been speculation concerning possible links between certain personality types (for example those who repress anger) and susceptibility to psychosomatic diseases.



4 Charles Darwin (1809–82), following his journey in HMS *Beagle*, developed a chronic and incapacitating illness, characterized by lassitude, palpitations, headaches, sleeplessness and tremulousness. There has been much speculation concerning the underlying cause of these symptoms. Analysts have seen them as psychosomatic – a psychological reaction to difficulties in Darwin's relationship with his father, who is usually portrayed as a stern and cold disciplinarian.

mental retardation. Some mental defects are the result of genetic diseases. Such conditions include craniostenosis, in which there is premature closing of the cranial sutures (where the bones of the skull are joined) and skull deformity and a number of serious metabolic disorders. Conditions such as mongolism or Down's syndrome are associated with aberrations of the number or shape of the chromosomes.

Uterine factors

Uterine factors in mental retardation include severe dietary deficiency in the pregnant mother giving rise to fetal damage. Virus infections, such as rubella (German measles), contracted by the mother early in pregnancy, can give rise to physical abnormalities and mental deficiency. Irradiation by X-rays or from an atomic explosion may damage the fetus as may certain drugs, such as thalidomide, taken by the mother during pregnancy. Rhesus factor incompatability of a severe nature is also associated with mental retardation in the newborn. Abnormal labour, with prolonged asphyxia and brain trauma during delivery, and premature birth may also result in intellectual deficiencies.

During infancy and early childhood serious infections, such as meningitis and poliomyelitis, and the rare central nervous system involvement in viral conditions such as mumps, measles and whooping cough, can result in serious mental retardation.

Prevention and treatment

The management of mental retardation can be divided into primary, secondary and tertiary prevention. Primary prevention includes genetic counselling and those medical measures, particularly in the field of obstetrics, aimed at reducing prenatal, natal and post-natal complications. Secondary prevention is achieved by the early identification and treatment of hereditary metabolic disorders. Tertiary prevention involves management of the mentally retarded so that the maximum potential of even the most seriously handicapped can be realized. For the mildly subnormal, special schooling may be required [Key]. For the more severely retarded, institutional care may be necessary.

KEY

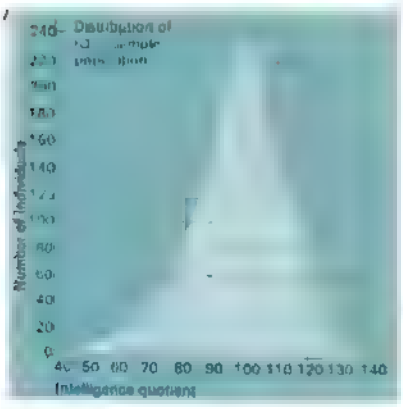
Mongolism, or Down's syndrome, is a chromosomal abnormality and occurs approximately once in every 800 live births. The overall appearance bears a superficial resemblance to that of members of the Mongol race. The head is small, the neck short and thick and the face has a flat-bridged nose. The eyes are slanted and the hands are broad with short stubby fingers. Mongols show severe mental retardation.

5 Alfred Binet (1857–1911), an experimental psychologist at the Sorbonne, Paris, devised a standardized intelligence test to discover mentally defective primary school children. The test was developed from intelligence scales for the investigation of normal and subnormal children. These were extended to become an intelligence test for the 3–15 age range from which modern tests developed.

5

6 Extremely withdrawn or moderately mentally handicapped children can be aided by using simple behavioural modification techniques. Normally, the child's solitary play is encouraged or 'reinforced' by the teacher's attention [A]. But in this instance he is only rewarded by the teacher's full attention when he joins the group [B]. As a check on the treatment, the original reinforcement is reinstated [C] which at once encourages the child to revert to his former behaviour. The new treatment is started again and the desired behaviour – the child playing in the group – is encouraged until it is firmly instilled [D]. In time the teacher can leave the group to its own devices [E] and the child will remain with it. Such behavioural methods are being extended.

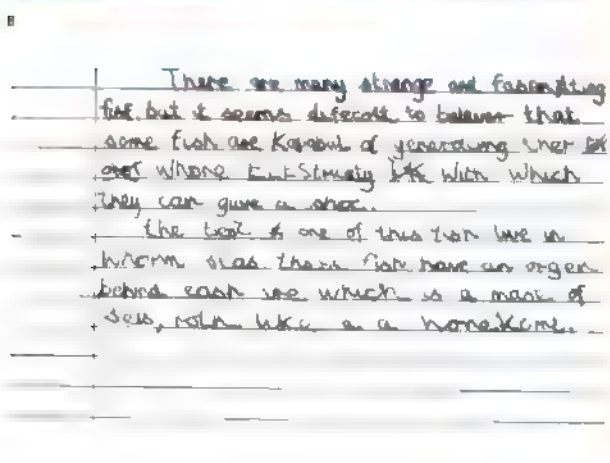
6 A



7 The distribution of adult intelligence quotients is from a sample of 2,052 people in the USA in 1958.

8 This example of dyslexic writing illustrates the bizarre character of the spelling disorders in dyslexia which may be connected with retardation although not always. Syllables may be omitted or put in the wrong order and sometimes it is not clear where one word ends and another begins.

8


There are many strange and fascinating fish but it seems difficult to believe that some fish use Kaboul of generating their own electricity which they can give a shock.
the boxfish one of this fish live in Indian seas. these fish have an organ behind each eye which is a mass of nerves, rather like a a homeykomb.

Physical cures for mental illness

Physical treatments in contemporary psychiatry include the use of tranquillizers, sedatives, stimulants and hallucinogenic drugs, electroconvulsive treatment (ECT) and certain psychosurgical procedures.

The use of drugs

A tranquillizer is a drug that induces sedation without loss of consciousness even when given in relatively large doses. The most potent tranquillizers are the phenothiazines of which chlorpromazine is the prototype. Synthesized in a French laboratory in 1950, this drug soon became widely used to reduce severe agitation or excitement of whatever cause. The phenothiazines appear to have selective effects on the delusions and hallucinatory phenomena seen in schizophrenia and manic depression. As a result they have become the drugs used in the treatment of these conditions. A long-acting phenothiazine, fluphenazine, has been developed that can be given by injection every 2–4 weeks and which maintains schizophrenics in a relatively stable state. The phenothiazines are also used to control excitable behaviour in manic states and in confusional states due to physical disease or drugs. They are not effective in the treatment of depression. Side-effects include drowsiness, muscle stiffness and rigidity, skin rashes, excess salivation and, occasionally, jaundice.

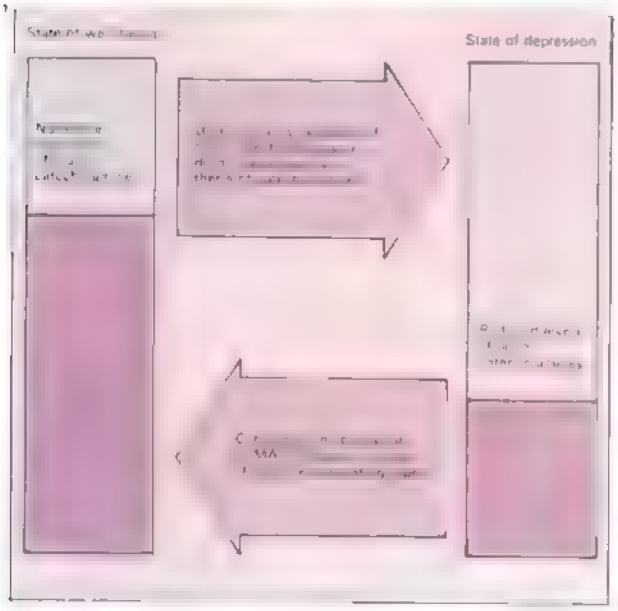
The most popular group of minor tranquillizers is the benzodiazepines, of which diazepam (Valium) and chlordiazepoxide (Librium) are the best known. These drugs, are used in the treatment of phobic anxiety, obsessive-compulsive disorders and minor anxiety and tension as well as in the management of the withdrawal symptoms occurring in barbituate and alcohol dependence. Another drug in this group, nitrazapam (Mogadon), is a popular and relatively safe hypnotic, used in the treatment of insomnia.

There are two main groups of antidepressants – the tricyclics, of which imipramine and amitriptyline are the most widely used, and the monoamine oxidase (MAO) inhibitors, of which phenelzine and tranylcypromine (Parnate) are the best known. The tricyclics compare quite favourably with ECT in the treatment of severe depression characterized by suicidal feelings, guilt, self-reproach, ideas of worthlessness, insomnia, weight loss and impairment of libido. The MAO inhibitors are less effective in such states but are thought to be useful in mixed anxiety-depressive states and in depressions believed to be reactions to some obvious environmental stress. They do, however, give unpleasant side-effects (palpitation, sweating and collapse) especially if the patient takes cheese, bean pods, some wines and other foods containing the substance known as tyramine. Lithium salts have been found to be useful in the control of mania, although why they work is not clear.

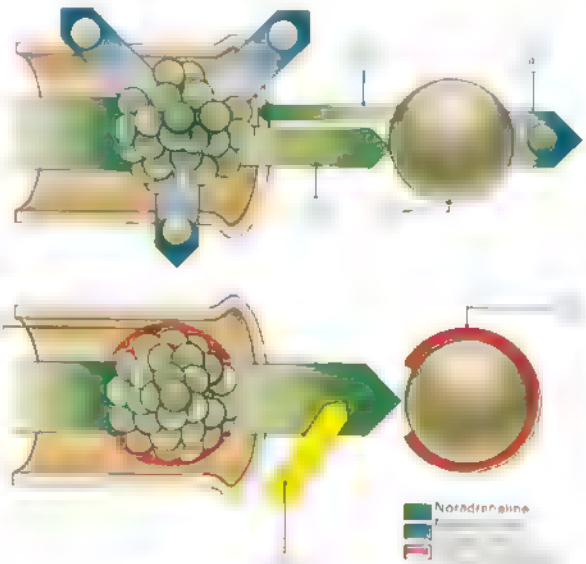
The powerful hypnotic drugs, the barbiturates, once widely used in the treatment of anxiety, have been superseded by the safer benzodiazepines. Because of their tendency to induce dependence, they are now used only as anti-convulsants in epilepsy. Similarly, the amphetamines, stimulant drugs once used to alleviate depression, have been replaced by the antidepressants and are now rarely prescribed. The hallucinogenic drugs, LSD and mescaline, have been used by some

CONNECTIONS

See also



1 The catecholamine theory of depression states that mood change is related to drug-induced changes in the level of chemicals called catecholamines in the brain. The theory is that a feeling of well-being is maintained by a continuous stimulation of certain receptors in the brain by catecholamines such as adrenaline, noradrenaline and dopamine. Drugs such as reserpine, a potent substance used in the treatment of high blood pressure deplete brain catecholamine stores and can cause lethargy, apathy and severe depression. In contrast, other compounds have been found that produce an increase in brain catecholamines and hence a marked improvement in mood.



2 A catecholamine neurotransmitter nor adrenaline is stored in a nerve ending. On stimulation of the nerve (A) noradrenaline is released (1) and exerts its physiological effect on the receptor site (2). Some noradrenaline is reabsorbed by the nerve ending (3) to be used again while the remainder is broken down (4). Antidepressants (B) are thought to act by preventing (5) the breakdown and reabsorption of noradrenaline. This increases the amount of free noradrenaline acting on the receptor site. They may also inhibit the breakdown by monoamine oxidase of stored noradrenaline (6). Or antidepressants increase the site's sensitivity to noradrenaline (7).

3 Those areas of the brain believed to be affected by antidepressants, stimulants, tranquillizers and sedatives are shown here. Antidepressant drugs are thought to affect catecholamines stored in the midbrain. Stimulant drugs act on the reticular activating system (RAS) and the hypothalamus. Tranquillizer drugs also act on the RAS and may suppress RAS response in the cortex. They are also thought to affect the limbic system. Sedative drugs such as barbiturates, which are used to induce sleep and as anticonvulsants in epilepsy, have a two-fold action on the brain. They act on the RAS and on the cerebral cortex (see 4) thereby suppressing brain activity.

psychiatrists in the management of alcoholism and certain neurotic states but the results are unconvincing.

Electroconvulsive therapy

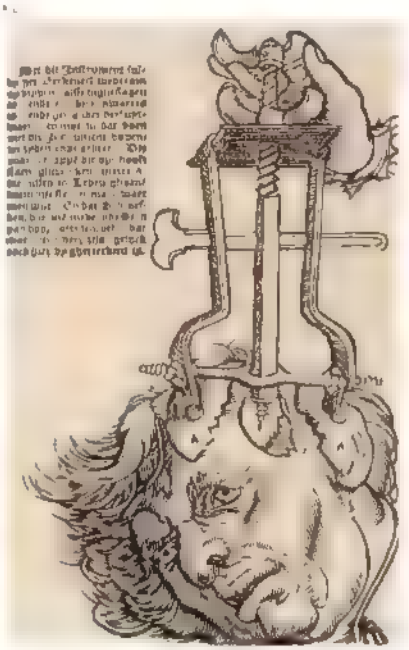
Electroconvulsive treatment was introduced into psychiatry in the late 1930s. It is still widely used in some hospitals and in severe suicidal depressions may be the treatment of choice. The usual course of ECT consists of 6-8 treatments spread over 3-4 weeks. The patient is given an anaesthetic and a muscle relaxant (which paralyses the main muscles and thereby reduces the severity of the convulsion). When unconscious, the patient has two electrodes applied to his head (one each side, in front of and above the ear). An electric current, usually 80 volts with a duration of 0.1-0.3 seconds, is passed between the electrodes producing an epileptic discharge and a convulsion, "modified" by the muscle relaxant and manifested by small twitchings of the facial, hand and feet muscles. The convulsions over, the anaesthetic and relaxant wear off and the patient regains consciousness. There is frequently an associated partial memory loss which is transient, but may take some weeks to clear completely. Studies have revealed that the crucial therapeutic element is the seizure discharge although why this exerts an antidepressive effect on mood is still not known.

Psychosurgery

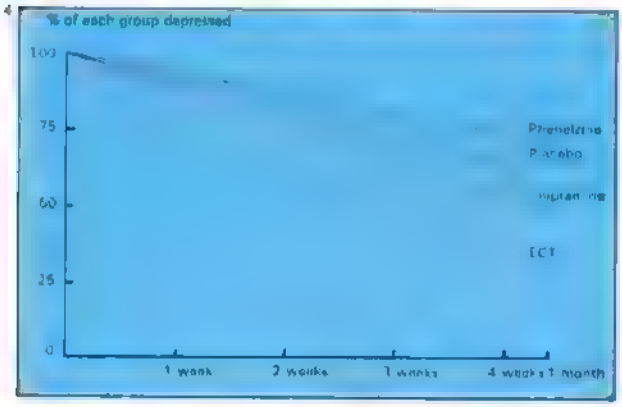
Psychosurgical operations involve the destruction or removal of normal or apparently normal brain tissue for the purpose of altering certain behaviour. Such procedures are among the most controversial in modern psychiatry. The first such operations were popularized by Egas Moniz [6] during the early 1940s and the so-called "standard" leucotomy (the cutting of certain selected brain fibres) was performed on thousands of chronically incapacitated patients.

The development of potent antipsychotic and antidepressant drugs during the 1950s resulted in a loss of interest in surgical procedures but the development of better operative techniques together with a renewal of interest in brain physiology has recently given new impetus to this field.

The primitive trephination procedures of medieval times were the forerunners of contemporary brain surgery, as a treatment for intractable psychiatric disorders. Historical discoveries have shown that decompression trephining for epilepsy and other brain disorders was common even among Neolithic Gauls and Bohemians, and often repeated five times upon the same person, the excised portions of the skull (rondelles) being used as amulets. In more recent times such procedures have been common among the Polynesians, the Aymaris of Bolivia and the Quichuas of Peru. Sharp flints were the instruments used.



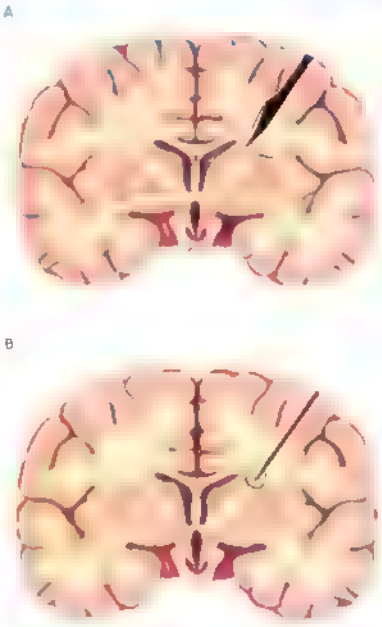
4 A large multicentre study of the relative effect veness of ECT, a tricyclic antidepressant (imipramine), an MAO inhibitor (phenelzine) and an inert placebo was carried out in Britain in 1965. The 250 depressed patients were divided into four treatment groups. Evaluation after one month showed that 71% of those treated with ECT had few or no symptoms compared with 52% for imipramine, 30% for phenelzine and 39% for placebo. This last figure may be taken to indicate the spontaneous short term remission to be expected in these depressions: they get better automatically.



5 Ugo Cerletti (1877–1963), the Italian psychiatrist, believed during the 1930s that epilepsy and schizophrenia rarely occurred in the same patient. He suggested that an epileptic convulsion might "protect" the patient against schizophrenia. Accordingly he induced epileptic seizures in schizophrenic patients, some of whom greatly improved. In fact, it is now known that the two are not antagonistic.

6 Egas Moniz (1874–1955), the Portuguese neuropsychiatrist, impressed by reports that the surgical removal of the frontal lobes in animals eliminated pathological behavioural responses to frustration, suggested that such a procedure might be effective in relieving anxiety in man. His enthusiastic advocacy of such therapeutic procedures led to their widespread adoption by many psychiatrists.

7 The technique of leucotomy, as performed in the early 1940s, consisted of making a burr hole in the side of the head above and in front of the ear. A cutting instrument – a leucotome or thin, sharp knife – was swept in an arc longitudinally, thereby cutting the connections between the frontal lobes and the rest of the brain [A]. The same procedure was then repeated on the opposite side of the head. With development of more precise surgical techniques and with modifications in the theoretical basis of psychosurgery, prefrontal leucotomy became less popular. Today, a probe is introduced through a burr hole [B] and is guided to the target area under X-ray control. The area is then destroyed either by the introduction of radioactive seeds through the probe or by the application of intense heat or cold. A favourite target [C] in chronic intractable depression is the inner aspect of the frontal lobe. Destruction of the anterior part of the cingulate gyrus is undertaken in obsessional disorders, of the amygdala in aggressive and hyperactive states and of centres of the hypothalamus in anxiety states and certain sexual disorders. A combination of centres may also be destroyed. Although Egas Moniz, the pioneer of psychosurgery, was awarded the Nobel Prize in 1949, there is still much dispute about the effectiveness of leucotomy.



Behaviour therapy

Many psychologists and psychiatrists, concerned with the causes of neuroses or mental illnesses, believe that they result from emotional problems which, once identified, can also point the way to the cure.

Behaviourist psychologists have a quite different approach. They argue that the cause of, say, a phobia about snakes is irrelevant and that the important thing is to get rid of the phobia by working on it directly. The individual learns his behaviour. It is therefore possible, the behaviourists argue, for him to unlearn or modify his behaviour.

Therapy by conditioning

Behaviour theory states that a neurotic reaction is acquired through a simple process of conditioning – the term used by the Russian psychologist Ivan Pavlov (1849–1936) in his experiments with dogs [1]. What is being conditioned is an emotional feeling of fear or anxiety. Such reactions easily give rise to more complex neuroses. A person with a conditioned fear of dirt may end up severely crippled by his compulsive need to wash off any dirt and with it his fear of contamination.

The behaviourist's answer to such faulty conditioning is to submit it to a process of deconditioning or counter-conditioning. One such method of deconditioning is extinction. In simple extinction the unconditioned stimulus is not repeated when the conditioned stimulus is presented. When applied to Pavlov's original experiments, for example, the bell is rung on a large number of occasions but no food is presented. Gradually the salivary secretion in response to the bell, built up in the earlier experiment, diminishes until it finally disappears completely. Simple extinction is probably responsible for the spontaneous "disappearances" of some neurotic illnesses.

One reason why all neuroses do not so respond may be that, unlike the Pavlovian dog, human beings can avoid things that frighten them and this avoidance, by relieving anxiety, serves as a further conditioning stimulus. For example, a woman with a fear of spiders will run away from any situation in which she may be exposed to these creatures. Her relief reinforces the conditioning process involved so that her phobia for spiders is strengthened. In such a case, behaviour therapy is required to force the individual to face up to the noxious stimulus.

Other behaviourist techniques

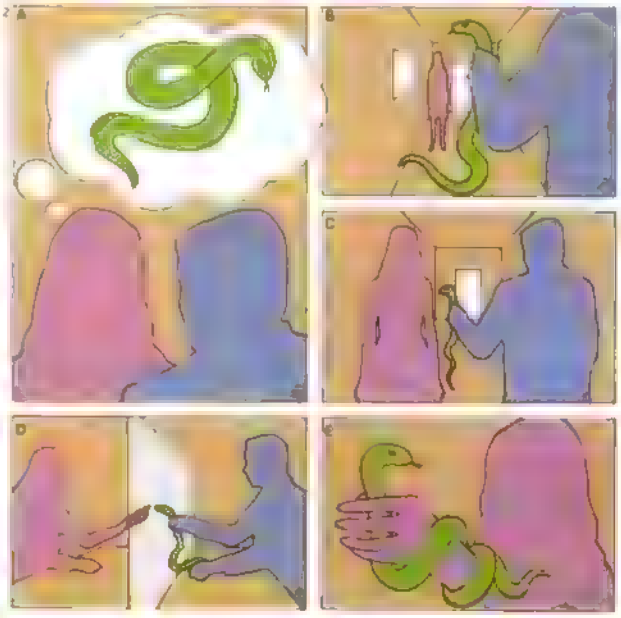
In systematic desensitization [2], an attempt is made to condition an alternative response to the fear-producing stimulus so that the occurrence of one inhibits the occurrence of the other. Anxiety involves the tensing of muscles. By systematically persuading the patient to relax his muscles while asking him to imagine the feared object (for example, a snake) in its least feared state, the therapist relieves the associated anxiety. Relaxation rather than anxiety is conditioned as a response to the imagined snake. Gradually the therapist works up a "hierarchy" of ever more threatening situations so that eventually the subject is able to tolerate the presence of snakes and even handle them.

Flooding [3] has been used with some success in the treatment of certain crippling neurotic states, particularly the severe obsessive-compulsive disorders. This involves making the patient experience his

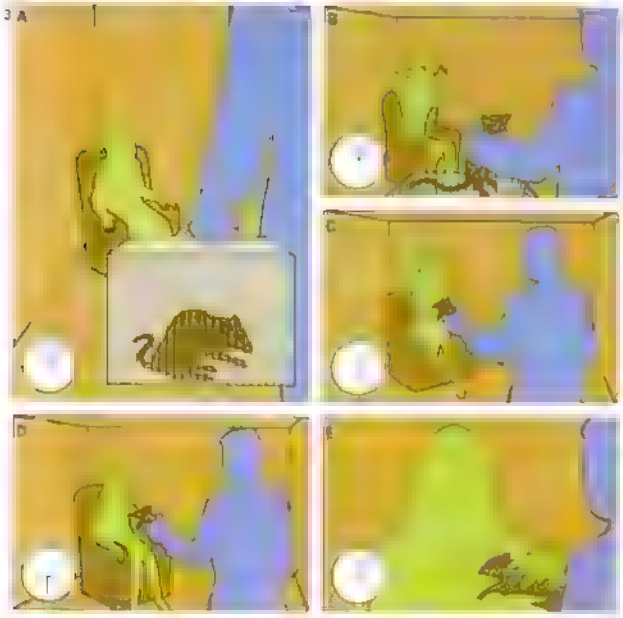
CONNECTIONS

See also

1 Ivan Pavlov, the Russian physiologist and Nobel prize winner, proposed the theory of the conditioned reflex as the basic model of mental activity. One of his experiments involved a bell being rung simultaneously with the presentation of food before a dog. Eventually the sound of the bell alone, in the absence of food, caused the dog to salivate. The dog had become conditioned by the experiment to the sound of the bell.



2 Systematic desensitization is a behavioural technique used clinically in treating phobias such as the fear of snakes. A reassuring therapist encourages the patient to think about snakes (A) until she can do so without anxiety. She is then confronted with a distant live snake (B) which is gradually brought closer to her (C). In the final stages of the treatment the patient progresses from touching the snake through a wire screen (D) to handling it by herself (E) without fear. Thus her fear response has been deconditioned and reconditioned into a fearless response brought about by positive reinforcements.



3 The aim of flooding or implosive therapy is to help a patient break the phobic cycle by having her make a deliberate effort to feel and experience her fear without avoiding it. The patient sits in a room while the therapist enters with the feared object, in this case a rat (A). The rat is taken out of the cage (B) and the patient screams with fear. She becomes progressively more frightened (C), but gradually she calms down due to the extinction of her severe anxiety reaction (D). Finally she is able to hold the rat (E). The crucial variable is the duration of exposure, for too early a termination of exposure to the rat increases rather than decreases the patient's great fear of the animal.

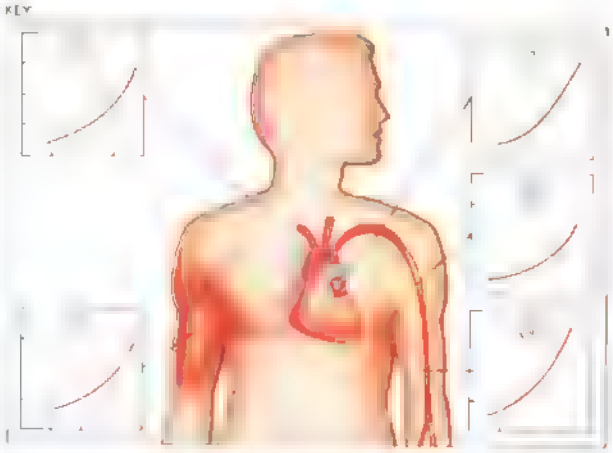
4 Unwanted behaviour can be diminished by giving unpleasant stimuli every time the behaviour occurs: so-called negative reinforcement or aversion therapy. A person's desire to stop smoking, for example (A), is reinforced by his doctor's warnings (B). Under treatment he is given an injection of apomorphine, a nausea-producing drug (C). Soon afterwards he begins to smoke in a contrived environment in which the "positive" aspects of smoking are over-emphasized but almost immediately he feels sick because of the drug (D). This process is repeated over several days. Later, in a normal social situation (E) the nausea returns whenever he is offered a cigarette (F). He soon stops smoking.

fear fully until it reaches a peak and then diminishes. Aversion therapy [4] has been used in the treatment of sexual deviations (such as exhibitionism and paedophilia) and alcoholism and drug dependence. The noxious stimulus may be a painful shock, or a nausea-producing or paralysing drug. Following such a course of aversion treatment, the patient is negatively conditioned

Operant conditioning

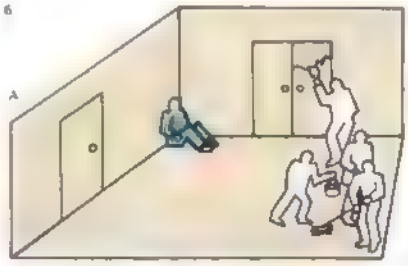
Operant conditioning, developed by B F Skinner [5] and employed in a number of treatment situations including so-called "token economies" [6], works by means of a mixture of rewards and punishments If a particular action is consistently followed by a reward ("positive" conditioning), then that particular action is likely to be repeated, if, however, it is followed by a punishment ("negative" conditioning), then the action is likely to cease. It is crucially important for the reward or the punishment to follow quickly upon the behaviour in question. Too long a gap results in a failure by the individual to connect the two and hence the conditioning quality of the reward or punishment is low

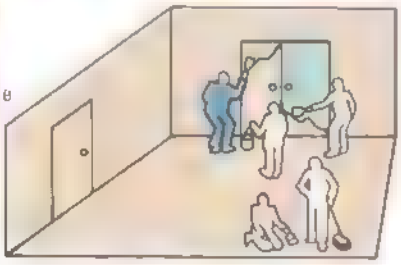
Behaviourists make considerable claims for their treatments. But while behaviour therapy has a significant role to play in the treatment of neuroses (such as simple phobias) in which there is a single fear-provoking stimulus to be deconditioned, its success rate in more complex states, particularly in the long term, has been questioned. In the treatment of personality disorders behavioural techniques have not to date been very successful, while for the psychoses they have little to offer In addition anxieties are often aroused concerning the ethics of certain forms of behaviour therapy, particularly aversion therapy It is felt that behavioural techniques are too much like "brainwashing" or the shaping of a person's character into something that it is not and that this is open to abuse Behaviourists insist, however that it is up to the individual to decide whether or not to receive treatment for his problem and that the ethical problems raised by such forms of treatment do not qualitatively differ from those raised by other forms of effective psychiatric treatment

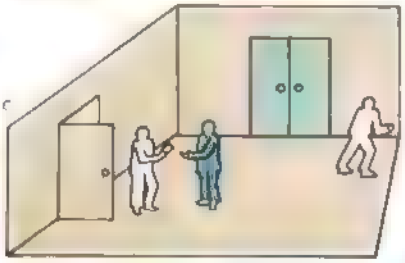


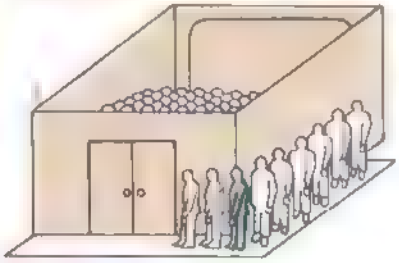
Many treatments of neuroses seek to reduce emotional arousal One way of assessing objectively the degree of stress or emotion experienced is to measure certain physiological changes during treatment Under stress the electrical resistance of the skin to the flow of an electric current decreases and the skin conductivity increases [1] The heart rate accelerates [2] and there is a constriction of the blood vessels causing a rise in blood pressure [3, Blood flow through the limbs increases [4] High levels of muscle tone during stress are measured by electrodes inserted in muscles [5]

5 The operant conditioning technique was first described by the American B F. Skinner (1904–) in the 1930s Whereas Pavlov took a hungry animal and paired a previously neutral stimulus (a bell) with the giving of food to an animal, Skinner waited for an animal to behave in a certain way and then stepped in with a reward such as food, so encouraging the animal to repeat the behaviour

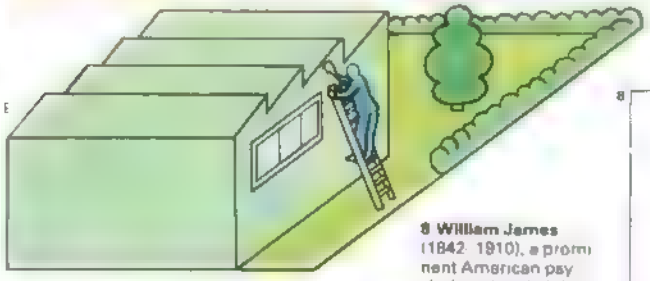




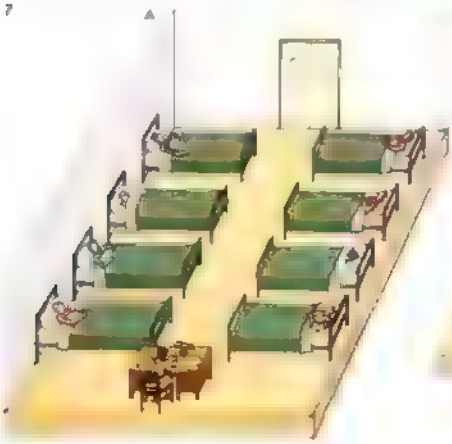


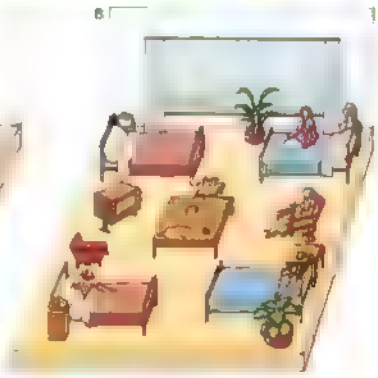


6 Token economies, a technique based on principles of behaviourist psychology are being used increasingly in the treatment of the mentally ill and handicapped Many such patients are socially withdrawn and do not participate in group activities [A] But whenever they do [B] therapists and nurses can reinforce such positive behaviour by rewarding them with tokens C These are equivalent to money within the hospital and are exchangeable for goods and privileges such as cinema or other entertainment [D] The patient's social behaviour can be gradually reshaped in this way and lead to a normal life [E] outside the hospital



7 The patient's immediate social environment can affect his mental state and behaviour, as psychiatrists have become aware in recent years As a consequence hospital wards today are no longer grim, large and impersonal with locked doors and few personal possessions [A] They have been redesigned so that they are smaller open and much more social with outside visitors being actively encouraged [B]





8 William James (1842 1910), a prominent American psychologist and philosopher and brother of the writer Henry James (1843 1916) laid the foundations of behaviourist psychology His theory that "emotion" is merely the perception of the bodily changes that accompany stress, which controversial has greatly influenced clinical psychology This attention on the outward symptoms of anxiety rather than the inward feelings provided an important theoretical framework for the research of the behaviourists

Psychotherapy

Psychotherapy is the general term for those forms of treatment in which conversation between the therapist and patient is the primary technique. The object of psychotherapy is not just the removal of symptoms, as with physical treatment of a physical disorder, but an exploration of the patient's personality. Indeed, in many cases the removal of symptoms ceases to be as important as the gaining of understanding and insight. To speak of "cure", therefore, as if psychotherapy were a surgical operation, is inappropriate.

Psychotherapy may be on an individual or group basis, may be interpretative, supportive or suggestive, and may be superficial or deep. By interpretative is meant the exploration by the therapist of something obscure or hidden to the patient. For instance, a psychotherapist interprets a dream by attributing to it some meaning over and above that given to it by the patient. Different psychotherapeutic traditions or schools, Freudian or Jungian for example, may give radically different interpretations of the same dream [1]. By supportive or suggestive psychotherapy is meant that the therapist adopts a reassuring or advisory approach. Deep psychotherapy refers to treatment which concentrates on the patient's early life or on repressed areas of his personality, while superficial psychotherapy centres on more recent and accessible experiences.

Patient and therapist

Not every patient is suitable for psychotherapy. Factors which favour a good response include youth, good intelligence, an ability to self-explore, a capacity to tolerate uncertainty, anxiety and frustration, and verbal skills. Psychotherapeutic techniques are most effectively applied in the treatment of the neuroses and personality disorders, sexual deviations and childhood disturbances. As well as the relatively orthodox analytically based group and individual psychotherapies, there is marital and family therapy, encounter group therapy, counselling, and the less orthodox "fringe" approaches which include Gestalt therapy, primal therapy, psychodrama and the Japanese Morita therapy.

Until recently, more attention was paid to the particular school of psychotherapy or the therapeutic approach of the therapist than to more personal characteristics. However, it has become clear that certain factors in the therapist's personality, such as ability to empathize, non-possessive warmth, sincerity and open-mindedness, are important.

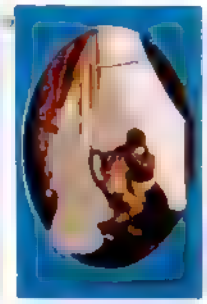
The most intensive form of psychotherapy is individual psychoanalysis. Psychoanalytic theory postulates that the individual defends himself against pain and anxiety by the use of certain mental mechanisms such as repression, denial and projection. These mechanisms are believed to be characteristically exploited in such neurotic states as hysteria and such psychotic conditions as paranoia. It is these mechanisms that are analysed and broken down in the psychoanalytic interaction.

Approaches to group psychotherapy

The development of group therapy has meant a greater availability of certain psychotherapeutic approaches. The more orthodox groups [2], which may meet daily or

CONNECTIONS

See also



1 Interpretations of dreams vary widely. A young man dreams [A] that he is climbing some stairs with his mother and sister. At the top he is told that his sister is pregnant. For Freud the dream reflected conflict between primitive instincts. This often has roots in a family situation [B]. The stairs represent sexual intercourse, the boy's incestuous wish for his mother [C]. The expected child is an indication of brother/sister sexuality [D]. Jung interpreted the same dream in terms of symbols or "universal archetypes" [E] that exist in all cultures. For Jung, the boy's guilt at neglecting his mother is revealed by his neglect of work (here symbolized by Dionysus in relaxed pose), his sister, a "love of womanhood" (here symbolized by Aphrodite), the stairs his passage through life (here symbolized by the Ages of Man), and the coming child his rebirth (here symbolized by spring).

2 Group psychotherapy is any form of therapy with more than two patients. Most groups consist of 6–10 selected patients. In group analytic psychotherapy, the therapist limits himself to interpreting the dynamics of the group. Verbal confrontation is encouraged but physical contact is minimal.

3 Encounter therapy is a group approach aimed at restoring spontaneity and involvement in social relationships by the creation of an atmosphere in which members can express their innermost feelings and explore new attitudes. Dramatic use is made of techniques such as body touching and mutual exploration.

weekly, are run under the leadership of a therapist trained in one or other of the established analytical schools. Less formal and orthodox groups have evolved over the years under the general label of "encounter therapy [3]. These groups tend to be less theoretical; they are relatively unstructured, geared to action rather than analysis and are mainly beneficial to those who do not suffer from severe personality disturbances or mental illness.

An offshoot of encounter therapy is marathon therapy in which the participants and therapists meet continuously for 24–72 hours. Enthusiasts claim that it is a particularly effective treatment for couples with marital problems and for certain neuroses.

Recent developments

Encounter groups often use principles developed by Jacob Moreno, the founder of psychodrama [4]. Janov's primal therapy [5] and Gestalt psychotherapy, with its emphasis on the patient's need to achieve behavioural change through understanding and experience, both involve intense, often exhausting participation and the acting-out of the individual's emotional turmoil.

Such treatments contrast sharply with a Japanese form of psychotherapy, Morita therapy, in which the patient, under the guidance of a specially trained mentor, undergoes a rigid and ritualized form of re-education that requires his participation in a number of prescribed social activities. Throughout his treatment, he is actively discouraged from baring his soul or analysing his feelings, activities that are believed to be at the core of all neuroses.

One therapeutic approach that has developed outside the main medical, analytical and behavioural traditions is counselling. Although concepts taken from medicine, behaviourist psychology and psychoanalysis are utilized, the major influence is derived from the American psychologist Carl Rogers and his client-centred therapy. Such an approach emphasizes the growth of an equal relationship between therapist and client, the mutual exploration of problems and possibilities and the rejection of any imposed system of attitudes or values.

KEY

The foundations of psychoanalytical theory and practice were laid by Sigmund Freud. Carl Jung, disenchanted with Freud's teachings, turned for inspiration to philosophy, religion and metaphysics. Alfred Adler emphasized the importance of man's inferiority feelings and need for power. Melanie Klein concentrated on the role of primitive infantile wishes. Jacob Moreno developed psychodrama and Carl Rogers advocated the adoption of non-interpretative psychotherapy.

4

5

6

7

4 Psychodrama combines both direct and evocative techniques. The patient acts out or watches others act out a personal problem. Other patients serve as both actors and audience. The therapist helps choose the problem, selects the actors, suggests the dialogue and guides the patients' general discussion.

5 Arthur Janov's primal therapy portrays neurosis as a defence against pain and seeks to induce cathartic emotional responses. Direct experience of pain is believed to be curative and so the patient is encouraged to undergo a series of "primals" (angry, fearful, sad or violent outbursts and upheavals).

6 Family therapy is a method of psychotherapeutic intervention requiring direct and continuous emotional involvement between family and therapist. For tactical or practical reasons some family members may be excluded from a few sessions but the focus and benefit is aimed at the whole family rather than individuals.

7 Marital therapy is designed to prevent marital disharmony and breakdown. In the typical joint interview husband and wife are treated together by two trained therapists, one of either sex. Their relationship is explored in depth and skills derived from psychoanalysis are employed in this kind of therapy.

Classifying personality types

Although each individual is unique and in some essential respects unlike any other person, people do, of course, resemble one another in such characteristics as physique, character, temperament and susceptibility to disease. Since time immemorial attempts have been made to classify people into groups on the basis of these resemblances.

The four humours and modern versions

For centuries the most generally accepted classification of human beings was the "humoral" theory propounded by the Greek physician Galen (c. AD 130–c. 200) [1]. According to Galen there were four different types of temperament – sanguine, phlegmatic, choleric and melancholic – each being produced by the preponderance of one or other of the fluids or humours that could be found in the body. If blood predominated the individual possessed a sanguine or optimistic temperament, if phlegm predominated, he was phlegmatic or unexcitable, if yellow bile predominated, the individual possessed a choleric temperament and was quick to anger, and if black bile predominated, he had a melancholic (nowadays, depressive) temperament.

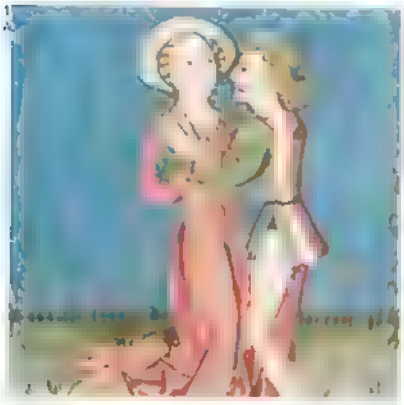
The humoral theory of temperament was generally believed in by both doctors and the general public until well into the nineteenth century and has left its mark in descriptive terms such as sanguine and choleric. Although no one today seriously credits that blood, phlegm and bile play any role in determining temperament, it has, nonetheless, had a curious revival in recent years. According to the British psychologist H. J. Eysenck (1916–), whose ideas are based on statistical research, human personality can be measured by two different yardsticks, one measuring stability-instability, the other measuring introversion-extraversion, the latter being terms made popular by the Swiss psychiatrist Carl Jung (1875–1961). Extraversion is the tendency to be 'outgoing' and interested in what is going on outside oneself; introversion is the tendency towards introspection and interest in one's own responses to outside events.

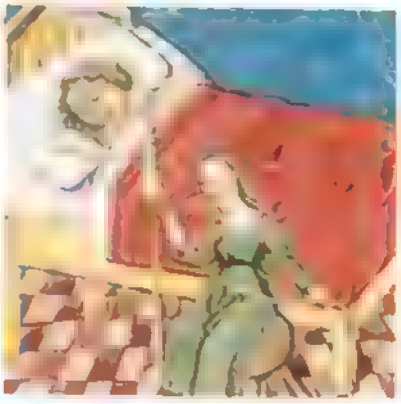
If both yardsticks are used [3, 4], four groups of people emerge – unstable introverts, unstable extraverts, stable introverts and stable extraverts. Here, says Eysenck, the four personality types correspond fairly precisely to Galen's four temperaments – unstable introverts resembling Galen's melancholics, unstable extraverts resembling his cholerics, stable introverts resembling his phlegmatics and stable extraverts being sanguine. Further research by Eysenck suggests that most neurotics are melancholic unstable introverts, that most criminals are choleric unstable extraverts and that most healthy citizens have natures that are either phlegmatic or sanguine.

The remarkable aspect of this system of classification is the way in which it bridges an apparently unbridgeable gap between classical and medieval ideas and modern scientific thought. Eysenck's classification differs, however, in that Galen asserted that one was melancholic, or phlegmatic, or choleric, or sanguine while Eysenck proposes that a person can be two, but not three of these at the same time, for instance, if one is marginally unstable, one is either melancholic-phlegmatic or choleric-sanguine, depending

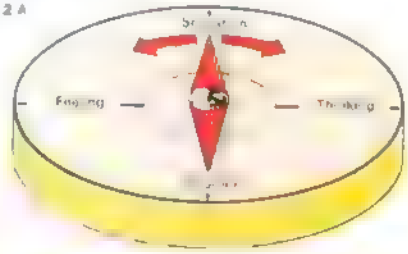
CONNECTIONS

See also

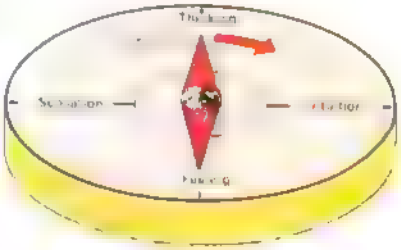


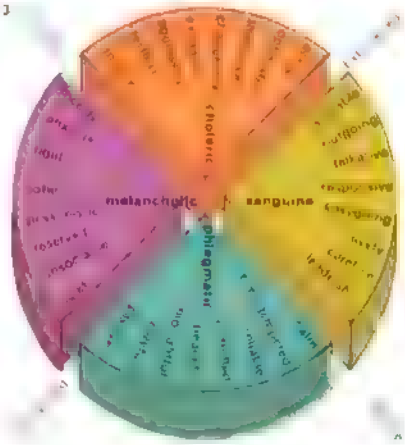


1 Character and temperament were for centuries thought to be the result of the four humours. These paintings present a medieval view of the sanguine [A], phlegmatic [B], choleric [C], and melancholic [D] man. Any imbalance between the humours leads to what are now called neuroses.

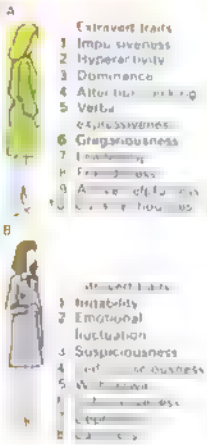


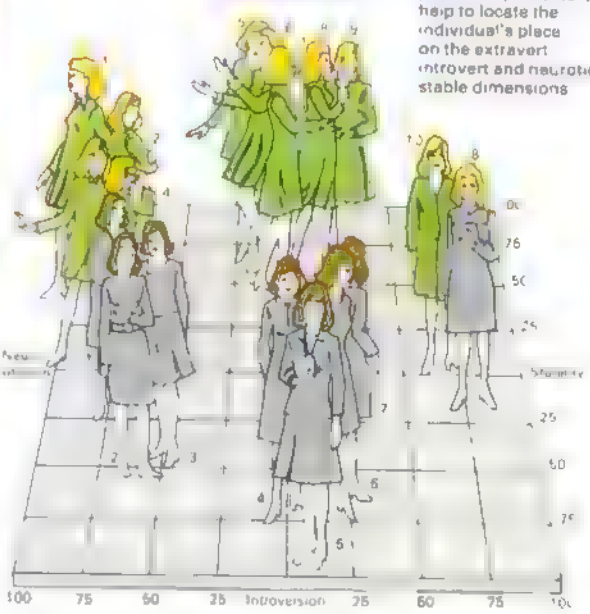
2 Carl Jung's compass of the psyche has four points. For a sensation type [A] intuition is least developed, for a thinking type [B] the feeling side is least in evidence (Feeling means the ability to evaluate experience without need to analyse.) As there is overlap in everyone one of the qualities on either side may be as highly developed as the dominant one.





3 According to H. J. Eysenck, every personality can be measured along two scales, one measuring introversion-extraversion, the other measuring stability-instability, and assigned a position in one of the four quadrants of a circle. This classification accords closely with the traditional four humours or temperaments, in which everyone is categorized as being either sanguine, phlegmatic, choleric or melancholic.





4 The answers given to a series of questions on an Eysenck personality inventory help to locate the individual's place on the extravert-introvert and neurotic-stable dimensions.

on whether one is introverted or extraverted.

Although Eysenck's theory of personality types is based on massive research, it is not universally accepted, largely owing to the fact that it leaves unresolved the question as to whether personality is innate and unchanging or the result of individual development and experience. For if the latter were so, it would be possible for a person's personality type to change over the years in response to changed circumstances, so reducing the value of assigning anyone to a particular type.

Other personality types

The Eysenck-Galen theory of personality and temperament is far from being the only one. Jung, in his book *Psychological Types*, divided people into eight groups, four groups of extraverts and four of introverts, the four divisions being in each case thinking, feeling, sensation and intuition [2]. Here also the classification has limited value since Jung held that we are unconsciously of the opposite type to that of our consciousness and that the fully integrated or "individuated" person transcends types. Freud was not a classifier but some of his followers have classified people according to the kind of mental illness their personality most resembles – and which they would presumably develop if they ever did fall mentally ill.

There is also an extensive medical literature that attempts to correlate physique, character and predisposition to both mental and physical illness, but its theories have as yet achieved little acceptance.

Reactions to classifications

People vary enormously in their emotional attitude to classifications of man since some people are fascinated by them and others find them morally offensive. Any classification that really worked would be of great practical value in medicine and vocational guidance. If it were proved, say, that tall, lightly built people were susceptible to one range of illnesses and were introverts, while short, heavily built people were prey to another range of illnesses and were extraverts – the German Ernst Kretschmer (1888–1964) did claim this [8] – the task of doctors and counsellors would be much easier than it is.

The Rorschach test, invented by Hermann Rorschach (1884–1922), invites a person to describe what he can see in ten standardised symmetrical ink blots. His answers being used as evidence of his fantasies, his personality, his psychological state and even his intelligence.

5 Thematic apperception tests are used particularly to measure personality improvement during psychiatric treatment. They comprise a series of 20 ambiguous pictures and drawings about which the subject has to make up dramatic stories. He is asked to indicate what the people are thinking and feeling, the events that have led up to the situation and the outcome. A man standing by a lamp post, for example, might be seen as waiting to meet his girl friend before they go out for a drink [A], or as just having had an argument with his fiancée and being shut out of her house [B]. The stories are then analysed, for it is assumed that the subject identifies negatively or positively with characters that he makes up and so will reveal something about his own inner impulse conflicts and sources of threat.

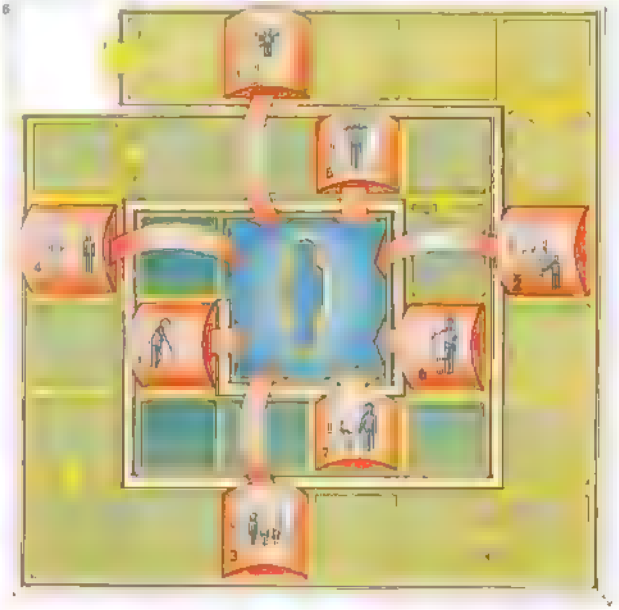
5 A

B

7 Thin, average-build and fat people, according to the American psychiatrist William H. Sheldon's theory of somatotypes, are endomorphs, mesomorphs and ectomorphs respectively. People tend to prefer objects that are similar in shape to themselves.

7

6



6 E. H. Erikson, the American psychiatrist, has theorized that we go through a series of psychosocial developments. Each stage is characterized by an identity crisis that strengthens the ego or self.

1. The development of hope in infancy.
2. The emergence of will and the determination to have independence and free choice with the exercise of self restraint.
3. A secure family gives purpose and initiative to pursue valued goals.
4. The development of competence casting aside childhood inferiority.
5. The adolescent develops fidelity and loyalty to the self and others.
6. The crisis of young adulthood – love – encourages mutual devotion.
7. Adulthood brings responsibility and teaches caring.
8. In old age wisdom grows.

8

8 Three distinct types of people, according to Kretschmer are the pycknics (broad, fat, short and tending to melancholy), asthenics (thin and tending to the schizoid) and athletics (muscular and intermediate in personality).

Thinking and understanding

Until the early 1950s it was generally believed that a child, at least for the first few months of life, was capable of only dimly appreciating physical sensations. Because a baby's sensory awareness and intellectual ability were supposed to be primitive, he was pictured as existing in a world that amounted to little more than a buzzing confusion.

The many talents of the infant

It now appears that within the first few weeks of life a baby has already learned to focus his eyes, see one image with both eyes and appreciate detail [1]. He can distinguish between different tastes and smells and between sounds of different frequency and loudness.

As a measure of the speed of this learning process, the baby, by the time he is 16 weeks old, is able to predict even the quite complicated trajectory of a ball that he has seen moving but which, during flight, disappears temporarily behind a screen. From birth the baby also actively selects what he looks at or listens to, attending to those things that differ in brightness, contrast, pattern or movement from objects he has seen before [Key].

Adults have an organized picture of the world that takes many things for granted. To arrive at this state involves a long, hard journey for the child. Perhaps the most influential researcher into this process is the Swiss psychologist Jean Piaget (1896–). From years of observation and inventive experiment with his own and other children, he has formulated the theory that cognitive or intellectual growth takes place in a series of four stages or phases. These he has called the sensorimotor [2A]; the preoperational [2B], the concrete operational [2C]; and the formal operational [2D].

During the sensorimotor stage the infant is constructing a picture of the physical realities of the world by touching, tasting, manipulating and destroying. By these means the building blocks for later thinking are built up. From then on each stage marks a growing ability to think in abstract terms that moves away from the physical reality of actual objects and involves reaching towards an adult grasp of the world. During the preoperational period the child has difficulty in understanding that an apparent change in an object is not an actual or real change. If two glasses of equal size are filled with liquid and then one is emptied into a taller but narrower glass, the child will maintain that the taller glass has more liquid in it. Later on the child will not be deceived by this, thus revealing an advance in thinking power towards the concrete operational stage.

Stimulating development

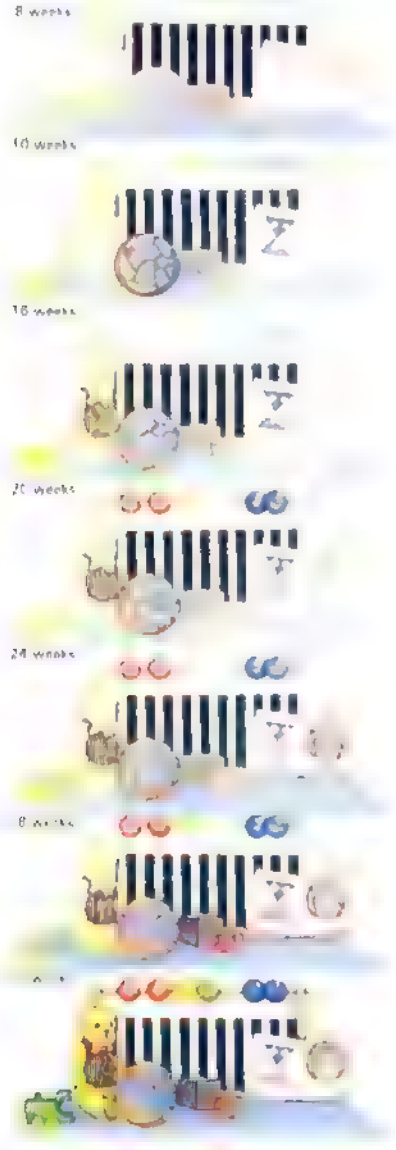
At the centre of Piaget's theory is the notion that what propels the child forward from stage to stage is "disequilibrium" – the situation that arises when a child's current picture of his world is shattered by the facts that his observations reveal. A young child will treat a magnet like any other toy, examining its size, weight and taste until he suddenly discovers that, unlike other objects, it attracts metal. This, according to Piaget, will put the child's world into disequilibrium and compel him to revise his picture of it.

A prominent feature of childhood is the markedly self-centred thinking typical of younger children. They assume that others see, experience and think about the world in

CONNECTIONS

See also

1 The acquisition of visual skills is an important part of intellectual growth. The ability to discriminate between lines, patterns, colours, sizes and shapes develops in a number of stages.



2 Jean Piaget, the Swiss psychologist, has shown that intellectual development, although a continuing process of discovery, can be divided into four main stages. In each stage different methods of understanding the world are used, each with its own logic and consistency, even if it does not conform to strict adult patterns. Piaget states that two processes, which he calls assimilation and accommodation, are essential for this growth. The child either assimilates new information into his existing view of the world, thus filling in more detail, or if this is not possible because it does not fit into his past experience, he accommodates the new information by revising his way of thinking. The increasing use by the child of accommodation indicates intellectual growth. For the first two years [A] the child is said to be in the sensorimotor stage of development. He understands the world through direct contact with objects and comes to realize that they have a distinct and separate existence. Gradually he learns that they continue to exist even when out of sight. He shows his intelligence through his actions. From two to seven years [B] the child is in the preoperational stage. Thought is more abstract and the image of an object is linked with thoughts about it. The child can also visualize something without seeing it but thinks everyone shares his view of the world. Unconscious contradictions are common. From seven to 12 years [C] the child is in the concrete operational stage. He is able to distinguish between present and historic time. He can order and arrange objects according to weight, size and volume and describe a series of actions (like walking to school) carrying the entire sequence in his mind. From 12 years onwards [D], the child is in the formal operational stage. Adolescents are able to use abstract rules to solve problems and they become more preoccupied with thought (and why they are thinking thoughts). They will perceive and explain things in increasingly objective, naturalistic terms. Presented with problems, they will consider solutions and choose the most suitable

3 Classical conditioning arises from "association". A child shows unlearned automatic fear of such things as loud noises. If a child sees a rabbit, about which he has no preformed ideas, at the same time as he hears a loud noise, he will begin to associate the rabbit with fear. Soon, the sight of the rabbit alone is enough to make him afraid. The Russian physiologist Ivan Pavlov was the first to observe and record such behaviour.

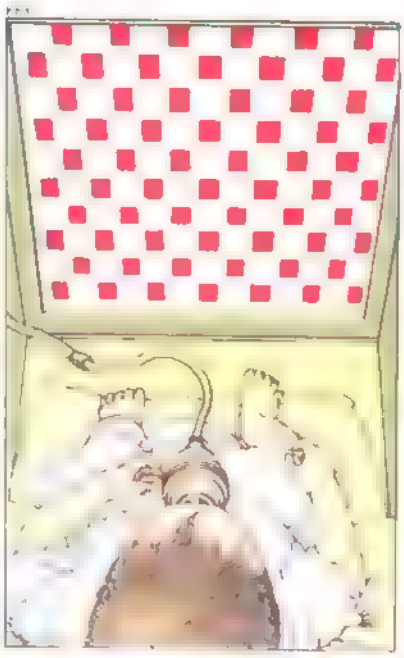
precisely the same way and from the same physical standpoint as they do. However, children are gradually made to realize that theirs is not necessarily the correct or only view. Another change in thinking that takes place during childhood is a decrease in attempts to provide magical explanations. Children more and more try to understand and describe their experiences in terms of physical cause and effect.

Learning involves a number of complex mechanisms. The Russian physiologist I. Pavlov (1849–1936), basing his theories on the results of his experiments with dogs, showed that we learn to connect impressions or bits of information by the process of association. He rang a bell at the same time as he fed the dogs and their mouths started to water. Soon the ringing of the bell by itself stimulated them. This is called classical conditioning [3] and many of our emotional reactions are learned on this principle.

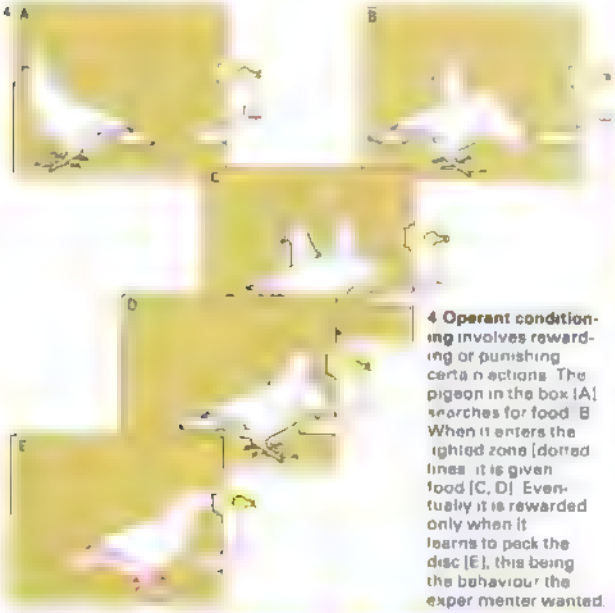
Operant conditioning

The American behaviourist B. F. Skinner (1904–) believes that we learn to do things or not to do things, by observing the reactions of those around us. We tend to repeat those actions that are rewarded and avoid those that are punished. Our behaviour is said to be positively or negatively reinforced. This is called operant conditioning [4] and is a decisive factor in moulding human behaviour.

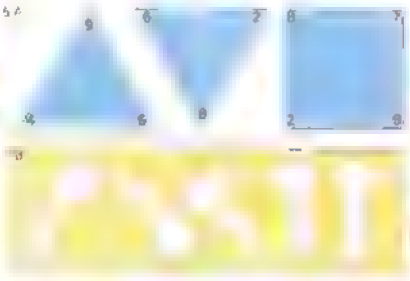
Operant conditioning has been found to exist even in newborn babies. In experiments, babies learned to perform quite complex tasks such as turning their heads twice to the right, three times to the left and then once to the right. The reward offered was the stimulus of a white light coming on for a few seconds. But once the babies had mastered the rules they lost interest in the experiment [6]. This reveals a crucial point about learning and reward: it is essential to find the right motivation before the child will respond positively and so extend his field of knowledge. The pre-school and elementary education system of Maria Montessori (1870–1952) is based on this self-directing drive to learn and on the belief that the child has creative potential and a right to be treated as an individual.

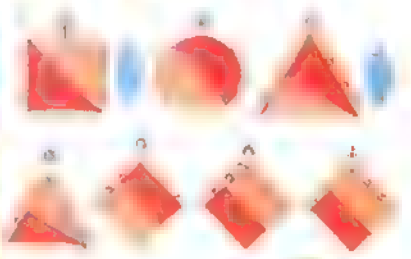


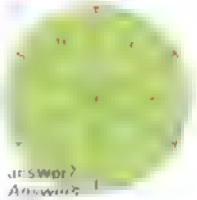
Recent technical advances in equipment and procedures have made it possible to carry out sophisticated studies even with week old babies. These involve the use of transistors, video tape recording, one way viewing screens and infra red scans. The apparatus shown here, for instance, tests the baby's ability to respond to a quite complicated learning test. By sucking on the teat he brings the picture on the screen into sharp focus. Results show that the baby soon learns to master the problem but requires the picture to be changed frequently to retain his interest. The findings of such studies have shown that young children have greater capabilities than had been thought.



4 Operant conditioning involves rewarding or punishing certain actions. The pigeon in the box (A) searches for food. B When it enters the lighted zone (dotted lines) it is given food (C, D). Eventually it is rewarded only when it learns to peck the disc (E), this being the behaviour the experimenter wanted.







5 IQ tests were first devised in 1905 by the Frenchman Alfred Binet (1857–1911) to select slow learners for special courses. Since then they have been used to measure almost everything of importance in intellectual development. Creative ability, productive thinking and problem solving were all assumed to be measurable in a single test. Such tests have recently become the subject of controversy on the grounds that they fail to measure intellectual growth, they do not show ability to learn and are often loaded in favour of a middle-class cultural approach. Attempts have been made to devise culture free tests that would avoid this bias. The four examples of "culture free" tests measure ability to see links between numbers and shapes (A) and (D) are numerical tests, (B) and (C) visiospatial tests.

Questions

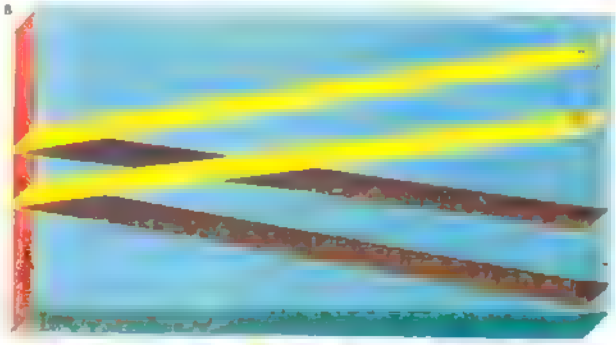
A and D. What is the missing number? B. Which is the odd one out? C. Which shape on the bottom row gives the correct answer?

Answers

A. The missing number is three. B. The odd one out is the fourth from the left. C. The pairing shape is the third from the left. D. the missing number is 15.

6 Habituation – or boredom – is one of the more important factors in discouraging learning and exploration. It is perhaps best illustrated by the common experience of a child soon getting bored with some new toy, such as a rattle. Children react positively to new challenges as long as the rewards for success are changed when they show signs of becoming habituated to them.

7 Chaining is the process of putting together related but separate movements into a coherent co-ordinated act. Many of the things that adults do without thinking – such as the seemingly simple opening of a door – in fact involve a complex series or chain of actions. The child has to learn these painstakingly by trial and error. Everything we do is similarly pieced together.



8 Environmental conditions can influence the IQ scores of initially intelligent children (A) and initially less intelligent (B). This difference may be accentuated (A1, B1) or even reversed (A2, B2). Many factors have been found to encourage high IQ scores. These include small family size [1], intelligent parents [2], who encourage the child [3], small well-equipped schools with high teacher pupil ratios [4], where the child, teachers and parents are on good terms [5] and a generally active life [6]. Conversely, either sort of child will develop slowly if his environment is poor or deficient.

Language development

The acquisition of language by the child is a remarkable achievement that has fascinated both parents and those who study child development. At present the central and fiercely debated issue of language development concerns the degree to which language structure and organization are innate or built into the human mind. At one extreme some researchers such as Noam Chomsky (1928–) maintain that every human being is born with both a natural capacity for language and an already programmed model or pattern in his brain. In contrast the "social learning" theorists, inspired by the work of B. F. Skinner (1904–), insist that cultural elements are crucial to language development.

The beginning of language

There is a close interplay between intellectual and language development. From the start the human infant is active and curious and by the first month of life he closely attends to the speech he hears. (The ability to hear is essential to speech development.) Although he can usually distinguish between different speech sounds much earlier, the child is usually a year old before he is able reliably to produce patterns of sounds that are identifiable as distinct words. Apparently the muscular control required for such a feat is gradually acquired, first through crying and then through cooing and babbling.

By the time he is about a year old the infant communicates his feelings and wants by varying the intonation, stress and frequency of his utterances. Soon afterwards, as he truly begins to become a speaking human being, he explores the way in which speech is organized and discovers the systems of rules (grammar) for putting words together to form sentences.

This breakthrough is dependent on the parallel development of the child's ability to perceive the world efficiently and to reason about it. He must be able, for example, to retain an abstract image of an object that he has seen, heard or touched before he can name it or respond to a name. Only in the second year can he hunt for a named object that is absent.

By the time he is about two years of age an infant can string together two or three words to form rudimentary sentences [2]. He can usually comprehend, as well as express, a variety of fundamental concepts about animate and inanimate objects and about actions and events. He uses such words as "here" and "there", for example, to indicate location and "all-gone" or "all-done" to indicate disappearance or cessation.

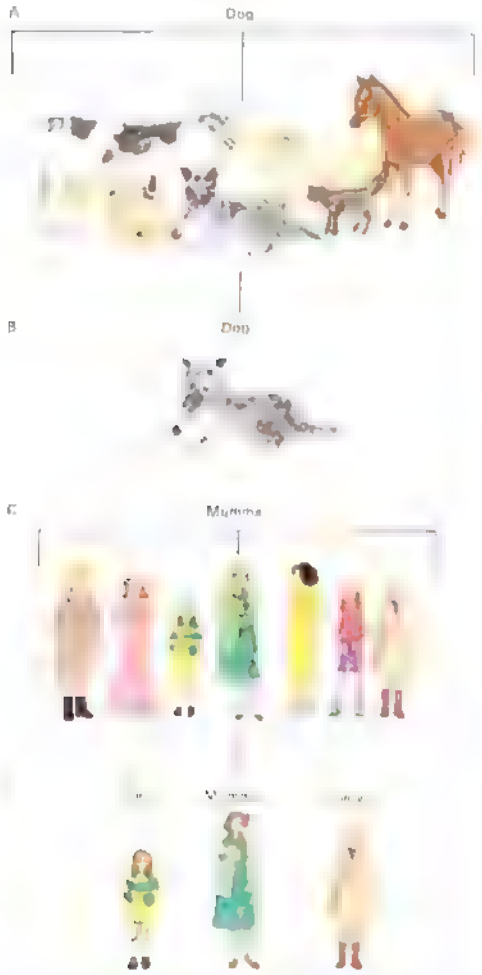
At this period, because the child responds positively to a request such as "Please fetch me that coat", parents are often misled into thinking that he understands more than he does. The fact is that conversation, particularly with children, is rich in non-verbal cues such as glances or gestures and it is only in combination with these cues that the child can make sense of the words he hears.

The growth of language

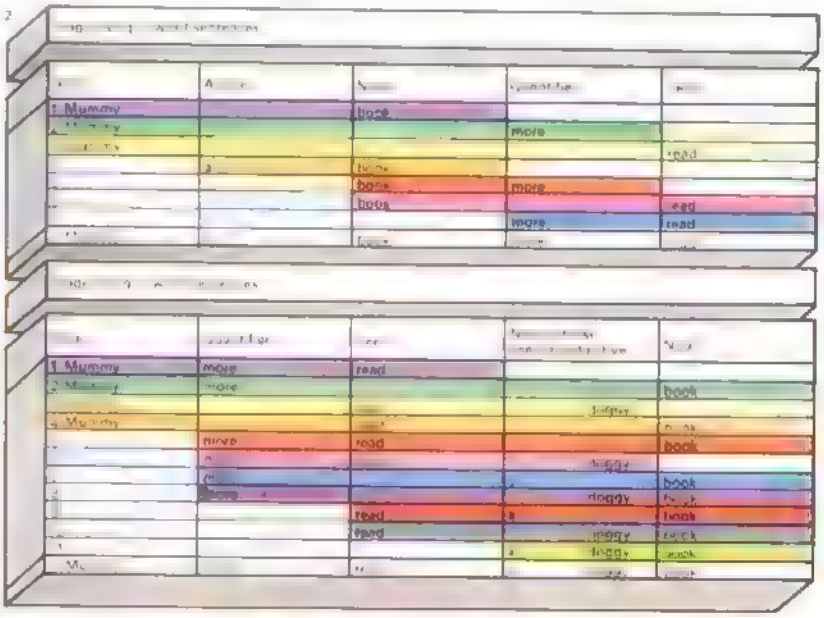
From about the beginning of the third year onwards the child rapidly becomes skilled at comprehending and producing sentences that do not depend on immediate context in order to be understood by others. By about the age of four or five almost all children have mastered the fundamentals of language.

CONNECTIONS

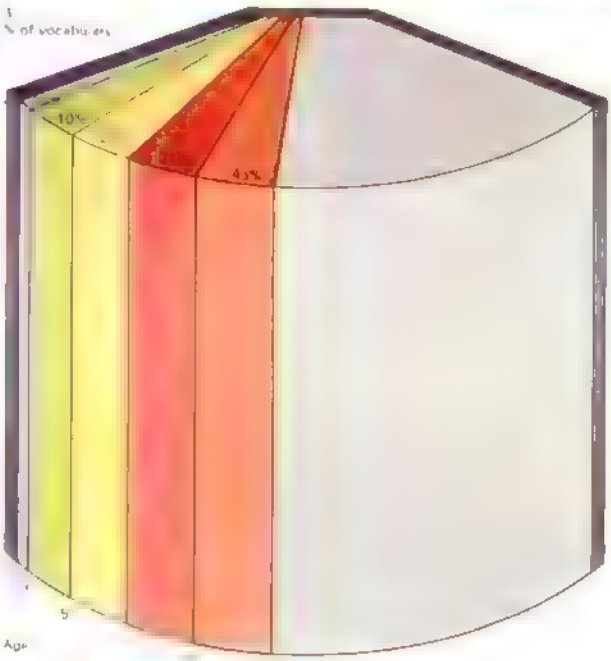
See also



1 Language development in children does not always occur at the same age but most children do show similar steps in language production and comprehension. Typically, a child's first words work as a complete utterance. The word "dog" (A) initially refers to all animals with which the child is familiar. Similarly, "mummy" (C) is used for all females with whom the child comes in contact. The word means "I see a dog" or "I want a dog" or some other animal or "I want mummy" or some other female. As the child develops, words gain more specific meanings. "Dog" (B) now refers to only one animal, a dog, and "mummy" refers to the child's mother alone and he uses other specific words to refer to all other females (D).



2 Children soon start to make up rudimentary sentences using first two words, then three. At the two word stage they choose between easy, unrelated words they hear and know are linked with what they want. They can then construct a number of sentences that express their needs. Once they are using three word sentences they are already using more complex noun and verb phrases and their range of expressions is wider.



3 The child's vocabulary can be seen as a percentage of his adult vocabulary. The child may use between 1 200 and 2 000 words when he starts school and be using between 3 000 and 5,000 words at school leaving age. Shakespeare used 29 000 words.

although they may have difficulties with certain sounds, such as the v in "Valentine"

One of the most interesting aspects of language development is undoubtedly that studies carried out in different cultures have shown the universality of the patterns and timings of word use already described. In addition, regardless of the specific language involved, children use similar, non-adult yet nonetheless formal rules in an attempt to impose order on their language almost as soon as they start to talk. They simplify language into telegraphic speech, producing characteristically over-regularized "errors", as in the plurals of nouns (mouses) and the past tenses of verbs (holded)

What parents can do

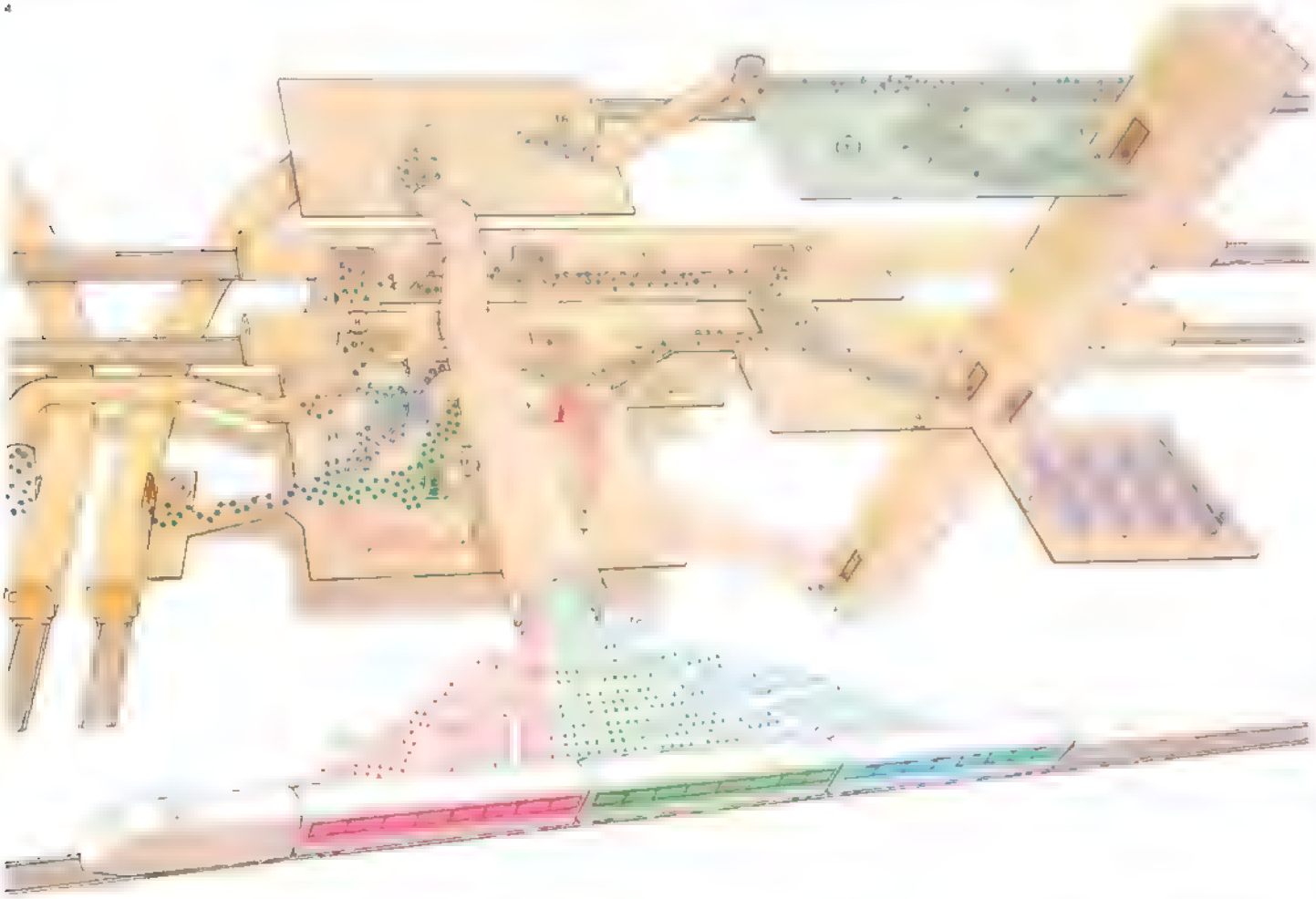
Language is not only an intellectual achievement but also largely a social product. Parents therefore play an important role in encouraging their children's interest in and practice of language by talking to them as often as possible, answering questions, describing events and by reading. In the early stages, when the infant cannot separate individual items from the whole, it is better to make connections between object and word as clear as possible. "Body" is less clear as a word than "hand" or "finger" since it includes so many different parts

Exactly what type of language is used between parents and children, whether normal adult or stylized "baby talk", does not seem to be very important and is more a question of current fashion than of objective scientific fact. It seems likely that for the first year simplicity and repetition are useful

One aspect of language development that often disturbs parents is difficulty in speech production. As many as one to two per cent of schoolchildren, for example, suffer from stuttering, mostly between the ages of six and ten. On average three times as many boys stutter as girls. There are various reasons why children stutter, both neurological and psychological. But whatever the cause it can in turn set up anxieties, especially at school where fellow pupils are quick to pick on and exploit a weakness, and this can set in motion a vicious circle. Should the problem persist it is best to seek professional help

A child's first words usually consist of only one or two syllables often repeated. Commonly they are only a part of a word, for example, "og" for dog. The parents' active co-operation and interest are an essential ingredient for language development

4



4 The child's rapid grasp of language is extraordinary in view of the complexities involved in making up a sentence. In this diagram the lexicon store [1] is the mind's memory bank for words. Language units assembled here [2] move down a lift [3] and travel to the deep structure, which contains the main rules for language usage. These are the syntax (grammar), the transformational rules and probably the semantics or meaning. From a departure area [4] monitored by an overall control centre [5] words enter a train [6] to journey around the language system. Between the departure point [7] and arrival point [8] words are given both meaning and syntax. Some words [9] may go on a further journey to acquire more information. Others, which have acquired a grammatical make-up [10], arrive in the transformational area [11]. Here they are sorted into connecting trains according to emphasis, active or passive [12], voice, negative [13] or affirmative moods and so on. Stray words [14] might cause an ungrammatical sentence if they managed [15] to enter the surface structure. Common words go straight from the lexicon store [16] to the surface structure [17], which governs the final presentation of words. Words emerge [18] ready to take their place in the sentence. Those that have fallen through the deep structure [19] emerge as ungrammatical language. The organized features of a sentence line up in the correct order [20]. A sentence [21] is then ready for use in the outside world

Social development

Being human is not just a matter of having an upright carriage, prehensile hands, a well-developed brain and organs that allow speech. These physiological and biological characteristics provide the "what" that makes up the human animal. But turning the "what" into a "who" requires our learning how to be human, and this is achieved through contact with our fellow men.

The wild boy of Aveyron

In 1799 a wild boy of about 11 years of age was captured in the forests of Aveyron in southern France by local villagers. He had been abandoned soon after birth, but had miraculously survived alone, living the life of a wild animal. His captors, however, found a human being only in the physiological sense. He could not speak, but grunted like an animal. He walked on all fours and seemed devoid of all the usual signs of human emotional responses such as pleasure or affection. A physician, Jean Itard, took the child into his care, determined to teach him some human characteristics. His success was limited. After several years the wild boy had learned only a few words and could just manage to eat properly and keep clean. Little more could be done, for he had missed the vital social and cultural process that psychologists today describe as socialization.

Stages in social development

Social development begins at birth and ends with death, but there is a crucial period in the child's early life during which the "social self" is formed. There seem to be four basic mechanisms that encourage all children in all cultures to become social beings: the desire to obtain regard, acceptance and recognition from others; the desire to identify with or be like those who are admired and loved; the fear of rejection and punishment; and the tendency to imitate. These four mechanisms influence different aspects of behaviour and dominate at different times.

The process begins at birth. The newborn infant has no specific values, attitudes or beliefs, just a propensity to develop into a social being. All theories of personality, regardless of their specific differences, stress the notion that we learn to be what we are because we interact with others, especially those who are accepting, supportive and encouraging, while guiding us according to those rules generally considered to be socially desirable.

In most cultures it is the parents who are the main socializing agents, with mothers being particularly influential because they normally spend more time with the child during the early years of development [1]. We are all affected by prevailing cultural standards, but the ambience of the family we belong to is a more specific and sharply focused influence because it filters and concentrates that culture.

The family's socio-economic class is also crucial because it conditions what is seen as desirable, probable and possible, and so how we come to regard the world – our aspirations, hopes and fears. The immediate depth of the family's influence is altered by other influences such as religious and political leaders and the mass media. The importance of these elements varies from time to time, but throughout, the original lessons learned at home are crucial.

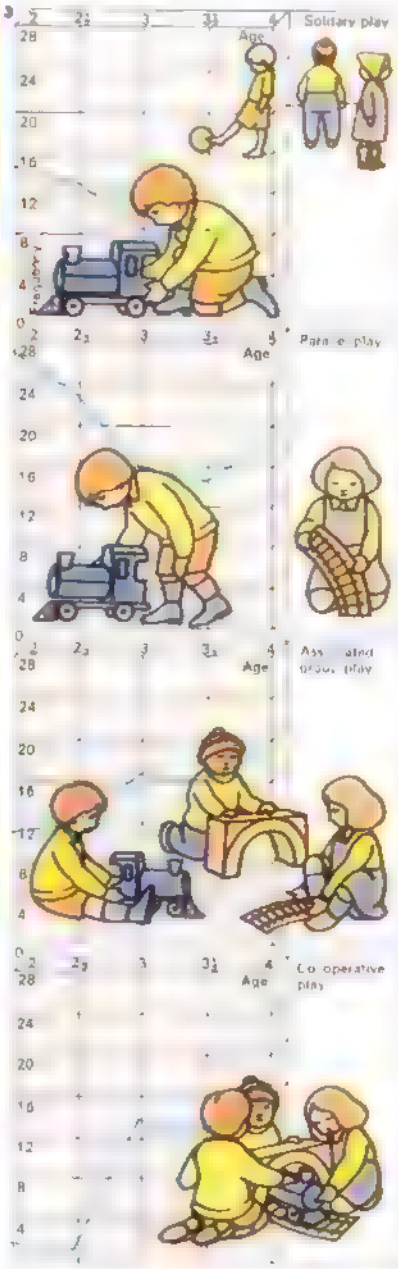
CONNECTIONS

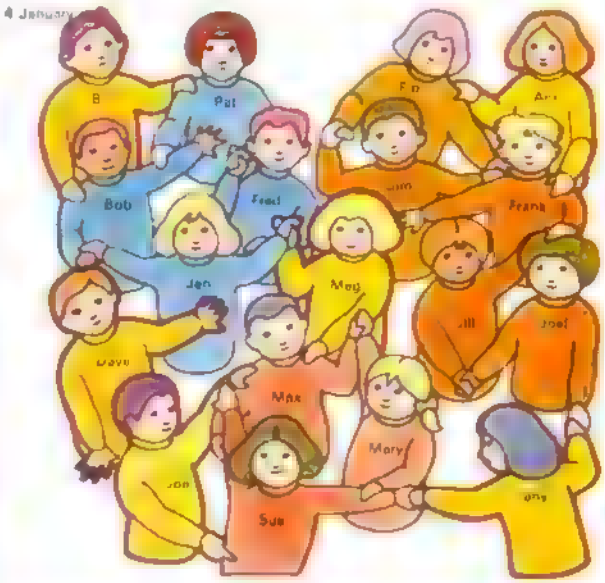
See also

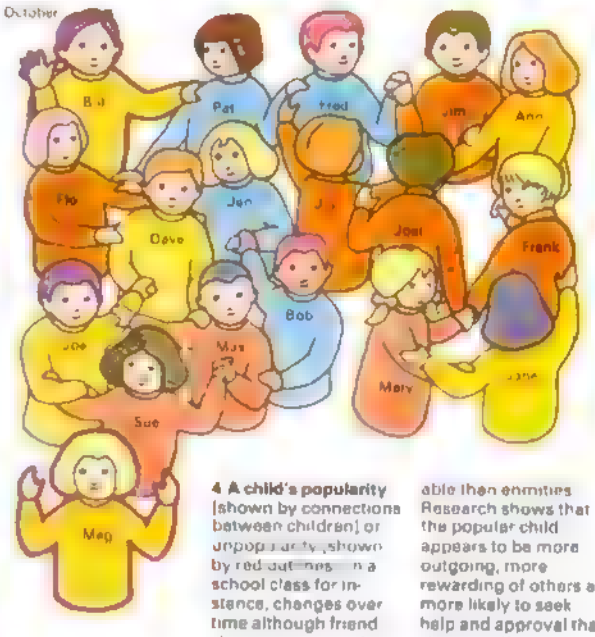
1 The child's early play during his first year primarily involves contact with his mother. He engages little in social play. In the second year, even when in close physical proximity with each other, children are usually quite content to play alone.

2 As children begin to meet more frequently, towards the end of their second year, play often ends in verbal or physical battles with each child demanding the toys and other objects or privileges for himself. This phase lasts for six months or more.

3 Different types of play serve a variety of purposes for the young child. Both solitary and parallel play involving little or no social activity and an apparent lack of effort to adapt to playmates, are prominent early in development and then decline between ages 2 and 3. At approximately three years, with increasing competence in intellectual, social and physical skills, children begin to seek out playmates actively. Their play takes on a more reciprocal and communal quality. In associated group and co-operative play children increasingly come to take into account the feelings, skills and responses of their playmates and others.







4 A child's popularity (shown by connections between children) or unpopularity (shown by red outlines) in a school class for instance, changes over time although friendships seem more durable than enmities. Research shows that the popular child appears to be more outgoing, more rewarding of others and more likely to seek help and approval than clinging affection.

Between the ages of two and five the child goes through a crash course of learning to make his impulses more social. He absorbs rules about property, the rights and welfare of others and the postponement of gratification of desires. The quality of that period and the way it is handled is critical. Affecting the family situation is the specific interaction between the parents' and the child's different temperaments, and those of his brothers and sisters if he has any.

After the child is five or six, when he begins to go to school, peer groups take on an increasingly important role in many areas of the child's development. Playing with peers provides the child with the opportunity to observe and then practise various skills.

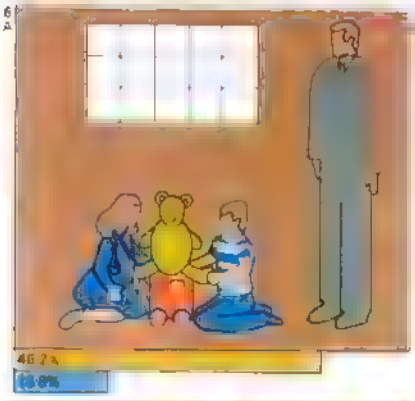
The importance of play is shown by its changing patterns [3]. Up to the age of 18 months or so solitary play is normal, and when children are together they tend to fight [2]. Then a change takes place and there is increasing contact with others in group games that involve and encourage friendly contact, reward-giving, the showing of affection and toy-sharing. That peers imitate behaviour has been shown in many experiments in classroom and laboratory [6]. If aggression is met by resistance it tends to diminish or at least shift focus. Recent studies have shown that when children of four or five see someone giving away toys during a game they are more likely to do the same than those who have not seen a similar model.

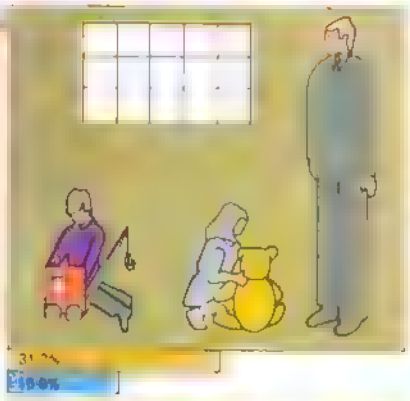
Learning sex roles

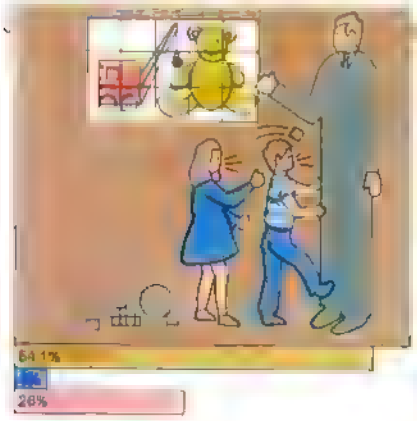
Biological differences set the stage for sex role development, and experiments have shown that from a few hours old boys and girls have markedly different patterns of gestures and grimaces during sleep. At the age of two-and-a-half the child is almost fully aware of his or her sexual identity. After this identity has been established it is difficult to reverse. To a varying degree parents convey their own concepts of what is appropriate sex-role behaviour – how one should relate to others if they are male or female and what kinds of work are undertaken by each [5]. Peer contact generally supports the cultural and sub-cultural attitudes originally conveyed by the child's parents.

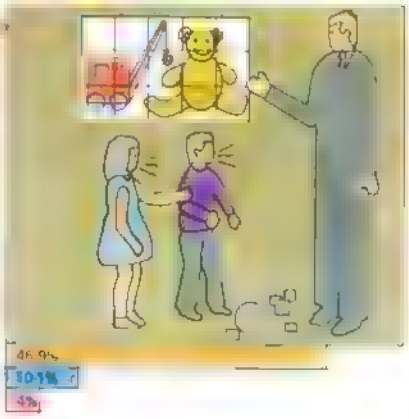
Adults may scoff at children's games of make-believe, but it has been shown that it is through such games that children explore themselves and others, testing different behaviours and attitudes. In games too, children learn the rules of society.

5 Boys and girls follow the unconscious behaviour patterns [A, B] presented by their parents almost from birth. In most Western societies males are encouraged to be achievement orientated, self reliant and less fearful generally. Girls are taught to be conforming, submissive and responsible. Increasingly, as stereotyped adult roles dissolve, children are exposed to more ambiguous influences [C].

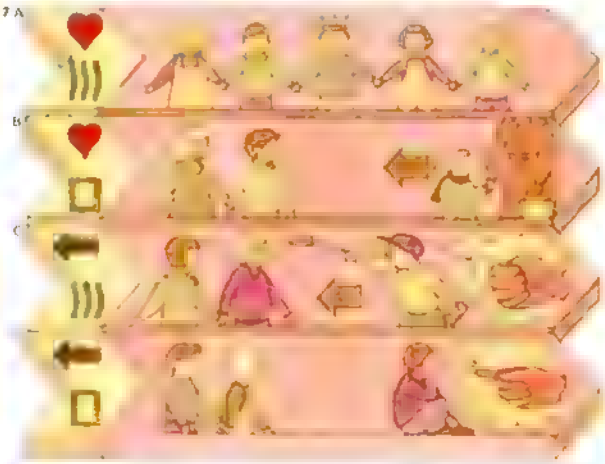








6 Studies of pre school children indicate that the more united the group initially, the more aggressively its members will be towards the outsider. Strong friends [A] and weak friends [B] were allowed to play with toys which were then taken away by the experimenter, and replaced with less attractive ones. In the resulting frustration the strong friends kicked the experimenter [C]. The weak friends merely called him names [D]. The proportions of time spent in inter friend co operation (brown), inter friend conflict (blue) and aggression to the experimenter (pink) appear below each picture.



7 Parental behaviour can be affectionate/permissive [A], affectionate/controlling [B], hostile/permissive [C], and hostile/controlling [D]. Each of these types is supposed to produce a different kind of personality in the child: [A], a friendly dominant child with high self esteem; [B] one more dependent, less friendly, less creative; [C] an aggressive child with poor self control; [D] a shy withdrawn child with a tendency to guilt feelings and insecurity.

Emotional development

From the moment of birth newborn babies differ widely in level of activity, degree of irritability and general responsiveness to their environment [5]. Today, emotional development – how people come to feel about themselves and others – is seen as being largely the product of early relationships and learning experiences both in the home and in the wider society.

A child's first emotions

The difficulty of gauging just what emotions the newborn child is experiencing is twofold. First, because he cannot communicate in words we have to rely on such non-verbal signs as frowns, grimaces and smiles, which are all open to misinterpretation. Second, our reading of the child's emotional state at any particular moment is strongly influenced by how we would expect to feel if we were in the same situation.

In 1917 the American psychologists J. J. B. Morgan and J. B. Watson proposed the theory that babies are born with three distinct and unlearned emotions – love, fear and rage – and that all others are refinements and elaborations of these. But it is now widely believed that there are no clear-cut distinctions in the emotional responses of newborn infants and that it is only after a month or so that their emotions can begin to be divided into positive and negative. Emotions such as elation, pride, anger and distrust gradually gain coherence over the next two years.

The evolution of the emotions is clearly shown in the child's response to strangers. Until he is about six months old the infant reacts positively to any friendly face. But, as his ability to differentiate between people increases, this indiscriminate response diminishes. By the time he is a year old the child will often react with considerable distress or fear to a stranger.

Fear, pleasure and first relationships

Intellectual development is significant in emotional development. Different fears, for instance, are characteristic of different ages [4], because to fear something implies the ability to hold it in the mind and be aware of its physical and emotional consequences.

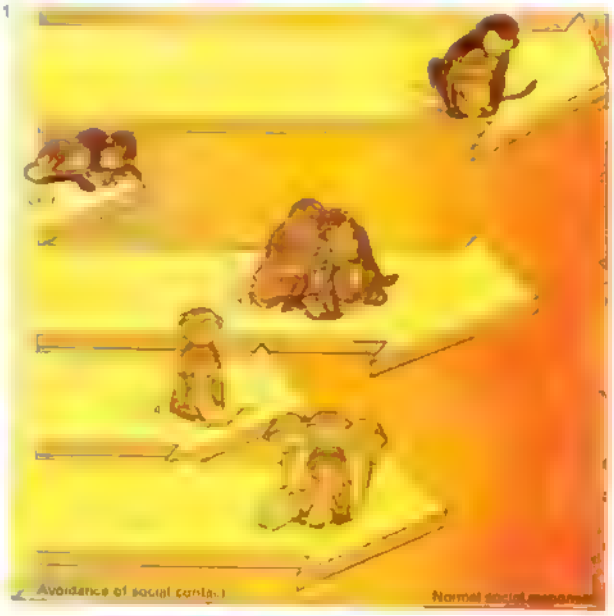
Jealousy, especially that prompted by the birth of another baby in the family, commonly surfaces at about 18 months and is at its most explosive for about two years after that. During this period the child may physically assault either the baby or the parents or both. With increasing maturity the outward signs of jealousy become much less direct.

Positive emotions have been much less extensively researched than negative ones possibly because they do not bring problems in their train. Pleasure is widely associated by psychologists with the gratification of instinctual drives such as hunger and thirst. But even the one- or two-month-old infant will show a positive reaction in situations in which the satisfaction of primary drives is not involved. He is receptive to friendly strangers, for example, and also appears to enjoy solving problems for their own sake.

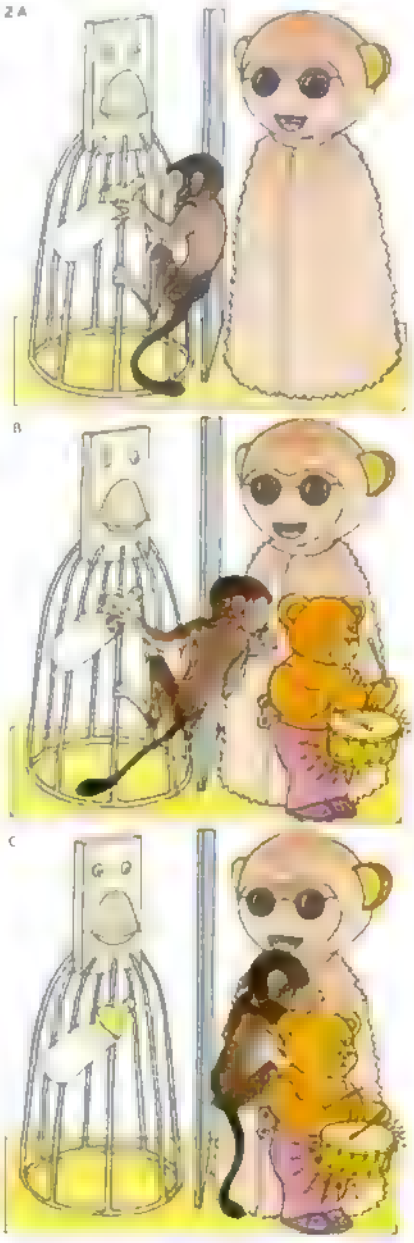
Recent studies with the young of the higher mammals, such as the studies carried out with monkeys in the 1950s by Professor Harry Harlow (1905–) and his associates at the University of Wisconsin [1, 2], and with human infants have focused attention on the complex relationship that develops between

CONNECTIONS

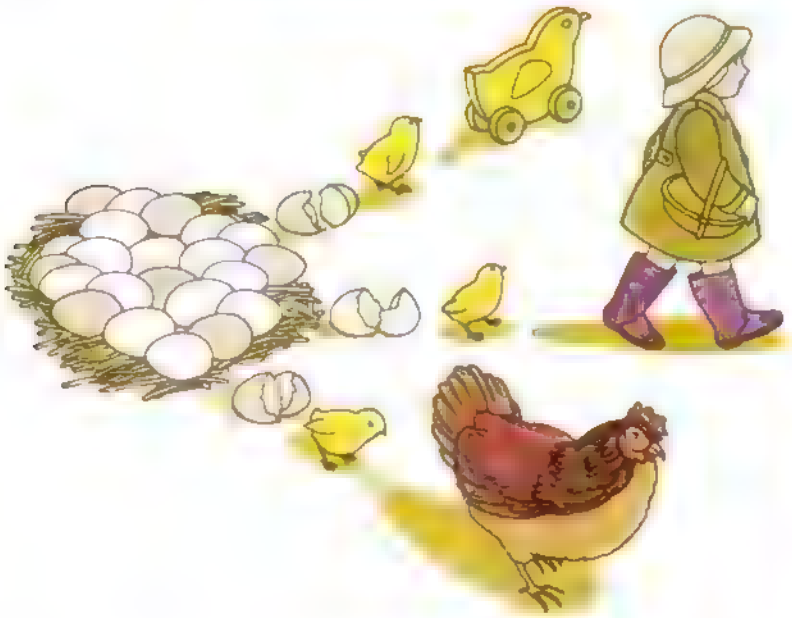
See also



1 The importance of family life in developing proper social responses has been shown in experiments with monkeys. Compared to a monkey raised from birth with its mother [1] one kept in total isolation [2], for more than the first six months avoids all contact and appears fearful, clutching himself and crouching. Monkeys brought up with siblings but without a mother [3] lead a normal life but indulge in more hugging than usual. A monkey raised with an imitation mother [4], although behaving normally with her does not show normal social or sexual behaviour on growing up. If the mother can move [5], the monkey is less fearful.



2 Comfort and shelter are by far the most important factors in the development of an affectionate social bond between mother and baby, as experiments with monkeys conducted by Harry Harlow demonstrate. Monkeys raised in cages separated from their natural mother but with both a wire mesh and a soft cloth covered surrogate mother spent more time clinging to the cloth mother even though it was the wire mother who provided the food [A]. When these monkeys were confronted with a strange object placed near them, like a noisy, drum beating toy bear, they reacted with great fear [B]. However, when the monkeys could cling to the cloth mother [C] this fear soon subsided. The cloth mother eventually came to provide a reasonably secure base from which to explore. The infant monkey would venture out to investigate novel objects but returned frequently for clinging and contact comfort.



3 The importance of the bond between mother and child is seen in the phenomenon of imprinting. After birth the young animal follows and forms an attachment for the first moving thing it sees. This is normally the mother, but any moving object, even a toy, can be imprinted on. The process starts even before birth, through sound. For if an egg that is hatching is placed near a loud speaker the chick will move towards that

the newborn infant and his primary caretaker (usually, but not always, the mother). The future emotional development of the infant is greatly influenced by this relationship. In animals this is also seen in the way that the young will form a permanent attachment (imprint) for the first moving object that they see, even if it is a human – as in the case of the geese that imprinted themselves on the Austrian researcher Konrad Lorenz (1903–) – and followed him everywhere [3].

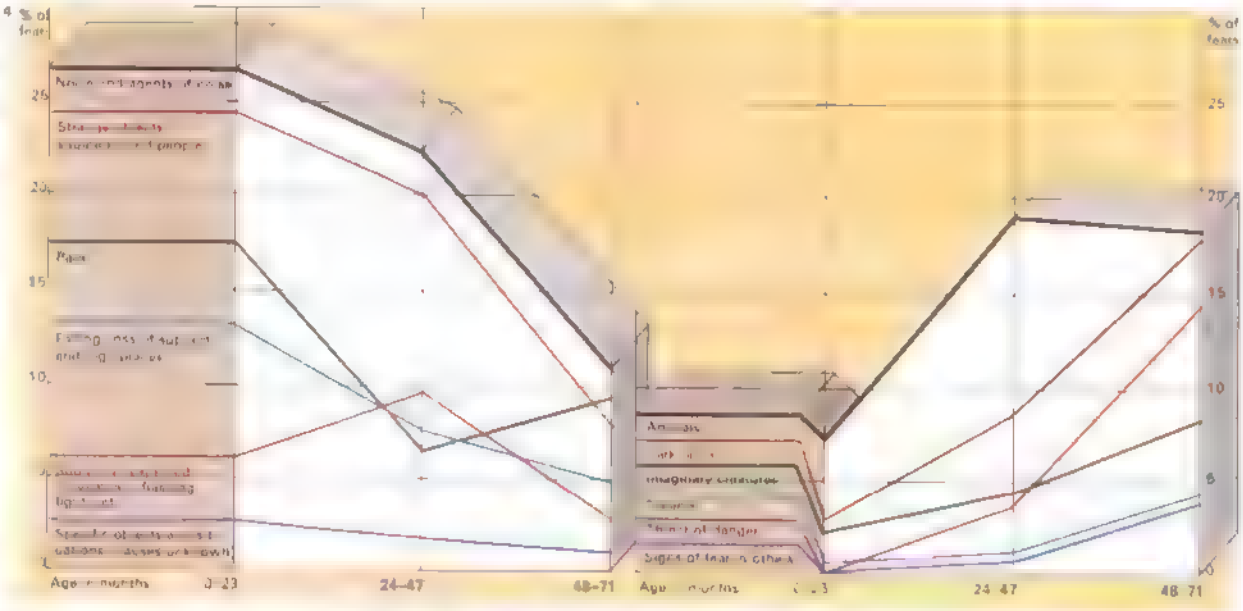
Between the ages of one and three the child is particularly vulnerable to emotional disturbance if the primary tie is disturbed for any length of time – as, for instance, when he is sent to hospital, where he is surrounded by a totally foreign environment. This can result in an extremely trying period of crying and clinging and, on occasion, even produces symptoms of depression and despair. Hospitals are gradually changing their once-rigid rules of excluding parents on the grounds that their presence would "disturb" the child. Where prolonged absences are unavoidable practical experience shows that they ought to be carefully explained rather than ignored and that the child should be surrounded, as far as possible, with familiar objects.

Erik Erikson, the American psychoanalyst, has stated that it is in his initial relationship that the child must first learn how to receive and give love, and move from complete dependence to increasing independence [6]. Erikson further notes that the child who experiences feelings of security and love in this primary relationship is more likely to develop into a person with a basic trust in the world and in his growing sense of self.

Learning to control emotion

The child must learn how to express and control the needs he feels. For, as with most areas of human behaviour, indiscriminate expression of the emotions is rarely permitted in the developing child. Aggression, a conspicuous feature of childhood behaviour, is an important example [7] of this. In infancy aggression is an immediate and shortlived response to the frustration of wants. But as children learn that violence is not acceptable, aggression and anger become less physically evident and more internalized.

The family environment is crucial in guiding the child's emotional development. A fearful child, for instance, is frequently an unconfident one who feels that he will not be able to meet successfully the challenges he is faced with. Stressing the child's limitations aggravates and prolongs the situation.



4 The nature of children's fears changes as they gain experience. Their emotional reactions to different situations and objects reflect developmental shifts in their understanding and knowledge. Children from four to six years of age are more likely than those below two years to be afraid of animals, bodily harm, dangerous situations, being alone in the dark, creatures of fancy or dreams. These fears depend on imagination and generalization from past experience, unlike those that are more likely to occur in younger children, such as a fear of sudden loud noises and strange situations, people and objects.

5 Marked differences in temperament or behavioural style are evident in babies only a few hours old. Such characteristics as tempo, adaptability, energy expenditure and mood, in existence from birth, have been found to persist relatively unchanged into adulthood. Where a child's innate style is compatible with the attitudes and expectations he meets in his environment, his development is likely to be healthy; where it is not there may be behaviour disturbances.

6 Independence is a sense of active mastery of the world, and starts as soon as the infant first discovers that he can influence his surroundings by twisting and turning. Later, crawling, walking and talking all add to his feeling of self-assertion. Dependence is the need for support in untried, uncertain or unexpected circumstances. Children tend to remain dependent if they are overprotected and not taught to follow some independent urges and activities.

7 Aggressive behaviour in children and factors influencing it have been studied by Arthur Bandura and his colleagues. In a series of experiments in 1963 they tried to determine whether live or filmed models acting aggressively towards an inflated rubber doll had a greater impact on the children's later expressions of aggression in similar situations. Regardless of whether the children saw live or filmed aggression they generally imitated the behaviour they had observed. They were also appreciably more aggressive than children with either a non-aggressive model or no model at all.

Moral development

Learning to judge the rightness or wrongness of conduct is crucial to a child's ability to adapt to society. But it is not easy either to give a universal definition of morality or to say how moral values are acquired. Some people have put forward the theory that there is an absolute morality to which everyone should aspire, whereas others adhere to the view that each culture sets its own rules and induces children to conform to them.

Inconsistent values

In the teachings of many religious and ethical systems, morality is sharply defined, right and wrong are presented as polar opposites and the assumption is made that a person who is moral tends to act consistently in accordance with high principles when dealing with other people. But a study conducted during the late 1920s in the United States by H. Hartshorne and M. A. May suggests that most people have flexible moral standards.

Everyone begins as an amoral infant who, as an active learner, acquires his first moral standards from his parents. These early standards are quite rigid and are tied to specific situations. The child often adopts them out of a wish to be obedient and avoid punishment or displeasure. The period between the ages of five and eleven is one of rapid moral development, with friends and teachers becoming significant influences on the standards of conduct adopted – largely through the effort to win their approval. With increasing maturity, obligations to society in general become relatively more important.

Although there are several ways of studying these changes in moral development, one of the most productive has been that of the American psychologist Lawrence Kohlberg. In face-to-face interviews, children are asked to respond to a number of moral dilemmas and the thinking behind their answers is then investigated [2]. Studies carried out in various societies indicate three basic levels in the development of moral reasoning. Kohlberg classified these as premoral, conventional and principled levels.

Steps in progress

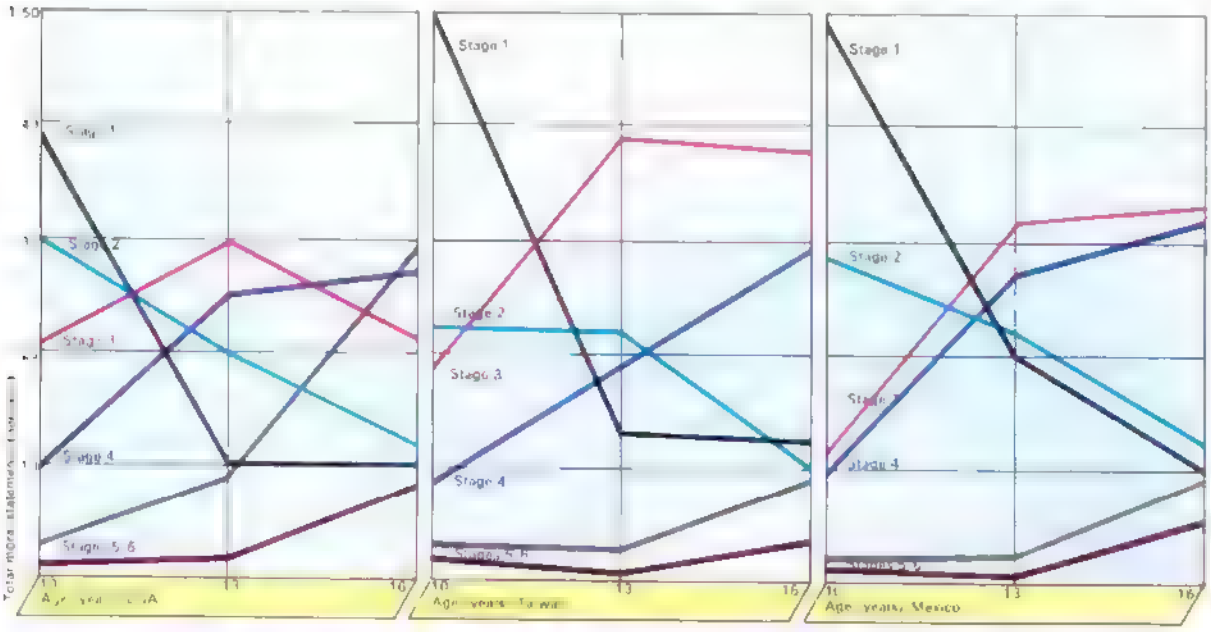
Each progressive step is considered to consist of a more complex and balanced way of understanding and thinking about the moral social world. Premoral reasoning is characterized by primary concern with one's own needs and self-interests. Conventional reasoning includes concern for others and an almost unquestioning acceptance of established authority. Principled reasoning takes into account the welfare of others and reflects self-chosen standards that are based on universal ethical principles. Law is respected but viewed as a human invention that is capable of being changed if circumstances alter.

Higher-level moral concepts are acquired only in late childhood or adolescence, apparently because they require an extensive foundation of intellectual growth and social experience. It appears that levels of moral reasoning may develop in a similar way in all children, regardless of their culture and their religious background [1].

The fact that an individual is able to reason on a principled level does not mean that he will always do so. Different moral situations evoke different responses drawn from all levels. Systems of morality may

CONNECTIONS

See also



1 Kohlberg's theory that moral reasoning develops in rising stages is shown in three different types of society. As the tested children grew older they provided fewer reasons for action at a premoral level (stages 1 and 2) and more at conventional (stages 3 and 4) and principled levels (stages 5 and 6). The graphs also indicate that moral reasoning seems to develop in a similar way in different cultures, since the pattern of changing values is basically the same in each of the three graphs. The tests were conducted among groups of middle-class boys of various ages living in urban areas of the United States, Taiwan and Mexico.

2 A child's level of morality is classified by Lawrence Kohlberg according to the response made to the following story. A sick woman is dying of a cancer from which only one drug can save her. A chemist sells the drug for 10 times what it cost him and the woman's husband has only half the asking price. He tells the chemist his wife is dying, but the price is not lowered. The desperate husband breaks in and steals the drug. Should he have done so and why? Answers both for and against are grouped in six stages. In classifying these it is not the choice of action that is important but the reasons that the individual gives for his action.

Premoral

Stage 1
If you let your wife die because you have not given her the drug, then you will get into trouble. Everyone will blame you for not spending the money to save her and there will be an investigation of both you and the chemist for your wife's death.

Contrary
You should not steal it, drug because the chemist will know that you were the thief and so you will be caught and sent to prison. And even if you do get away with it your conscience will bother you and you will always be worrying that the police will catch up with you.

Stage 2
If you do happen to get caught, you can always give the drug back and then you will not get much of a sentence. It will not be too great a hardship for you to serve a short jail sentence and it will be worthwhile as long as your wife is still alive.

Contrary
You may not get much of a jail sentence if you steal the drug but your wife will probably die anyway before you get out of jail and so it will not do you much good in the end. If your wife dies you should not blame yourself because it was not your fault that she got cancer.

Conventional

Stage 3
No one will really think that you are bad if you steal the drug, but your family will certainly blame you and think that you are an inhuman husband if you do not. If you let your wife die, you will never be able to look anybody in the face again.

Contrary
It is not just the chemist who will think that you are a criminal, everyone else will as well. After you steal the drug, you will feel bad when you think about the dishonour that you have brought upon yourself and your family. You will not be able to face anyone again.

Stage 4
If you have any sense of honour you will not let your wife die just because you are afraid to do the only thing that can possibly save her. You will always feel guilty that you were the cause of her death if you do not do your duty by her, whatever the consequence for you.

Contrary
You are desperate and you may not know that you are doing wrong when you steal the drug. But you will know that you did wrong after you are punished and sent to jail. You will always feel guilty about your dishonesty and the fact that you have broken the law against stealing.

Principled

Stage 5
You will lose other people's respect, not gain it, if you do not steal the drug. If you let your wife die, it would be because you were afraid, not because you had worked out that it was a wrong act. So you would lose your self respect as well as that of others.

Contrary
You would certainly lose your standing and the respect of others in your community as well as violating the law. If you stole the drug, you would also lose respect for yourself if you allowed yourself to be carried away by emotion and forgot the long range point of view.

Stage 6
If you do not steal the drug and thus let your wife die, you would always condemn yourself for it. No one would ever think of blaming you and you would have lived up to the letter of the law, but you would not have lived up to your own standards of conscience.

Contrary
If you stole the drug, you probably would not be blamed by other people, but all the same you would condemn yourself because you would not have lived up to your own conscience and the standards of honesty by which you have always lived in the past.

reflect several levels of reasoning. Traditional Christianity, for instance, preaches right actions both because they are in themselves ethical and because the jaws of hell yawn for those who transgress (exemplifying both principled and premoral levels)

Formal moral values

As children develop their own moral reasoning, the way in which parents and friends react to their ideas becomes particularly important. Parents who take seriously their children's opinions on moral issues and discuss them are more likely to have children who reason at a conventional level than at the premoral level.

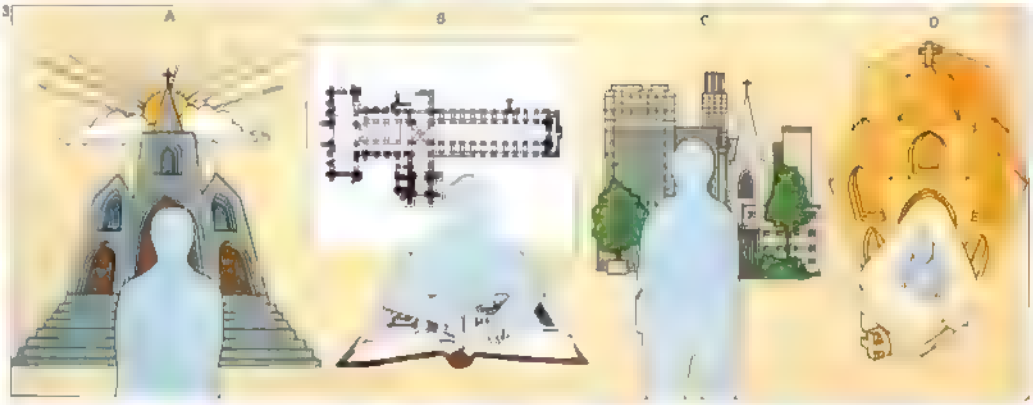
What parents, friends and others who serve as models actually do has a powerful influence. While words are likely to influence the development of a child's moral reasoning, they do not necessarily result in changes in moral conduct. In general, children copy what others do rather than what they say and are likely to develop consistent moral values only if the actions of influential people in their lives do not conflict with their words.

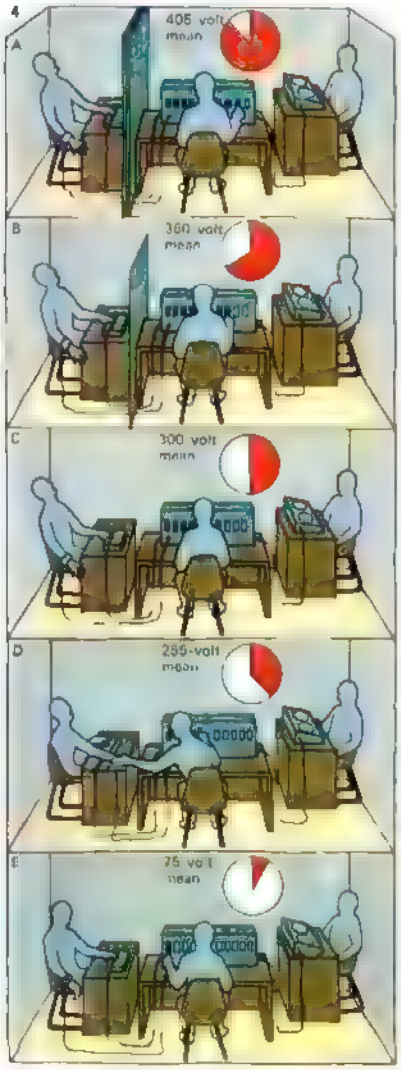
How a child feels about his own thoughts and deeds plays a pivotal part in his moral development. Typically, a sense of guilt and self-regulation develops gradually and so has an increasingly important role in monitoring behaviour [5]. As this occurs the child starts to regulate his own moral conduct. The child behaves morally principally to avoid his own feelings of guilt.

One of the clearest ways in which parents and others contribute to the development of a child's self-control or sense of guilt is through discipline for disapproved acts. Some research indicates that the most effective discipline is a brief withdrawal of affection combined with a detailed explanation. The latter encourages the child to see things from the perspective of another person and promotes the ability to form independent standards. The timing of discipline may be decisive. It is likely to be most effective if it occurs just before or just as the child begins to do something that is forbidden. When this is not possible, it is helpful if the situation can be carefully re-created by talking with the child and describing the action.

The American researchers Hartshorne and May tested the moral values of thousands of children during the 1920s by means of games and contests in a wide range of situations. They found that if a child cheated in one situation it was not possible to predict whether or not the same child would do so in another situation. They showed too that a child who cheated on one occasion was not necessarily the child who lied then; nor was a child who lied to an adult necessarily the child who lied to a friend. The question in the end is probably not whether children or adults will behave morally or immorally, but rather in what situations they will do so.

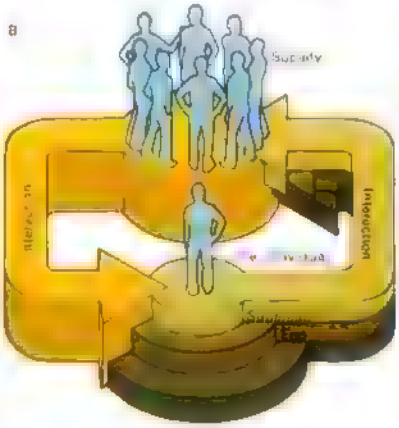
3 Our perception of the world is radically affected by our moral values. The sight of a church, for instance, will evoke quite different responses in different people. To a priest [A], it is the sacred house of God, while to an architect [B] it may be merely a construction. An atheist [C] may not distinguish it from surrounding buildings, yet to a Marxist [D] it may represent oppression.





4 Individual scruples and the demands of authority are often at odds. A soldier in uniform will kill but the same man in civilian dress might not. To test response to authority, an American, Stanley Milgram, devised a scheme in which volunteers took part in an 'experiment' requiring them to give increasingly severe electric shocks (from 15 to 450 volts) to a protesting man (really an actor) to correct mistakes in a learning test. Authority, in the guise of a "scientist" told volunteers they must go on despite a rising crescendo of protest. Where the 'learner' was remote [A], 25 out of 40 people went to 450 volts, the mean being 405. Willingness to obey orders declined as the "learner" was brought closer [B, C, D] but it was not until the choice of maximum shock was left to the volunteer that the level fell to 75 volts [E]. Control groups predicted they would disobey rather than cause pain.

5 The development of a conscience is explained by behaviour(ist) psychologists as the outcome of rewards and punishments [A]. A child starts with no model of how to behave [1]. When he commits a forbidden action [2], the outside world punishes him [3] by at least withdrawing support. Thereafter, in theory, he modifies his behaviour to avoid the mental scar of the action and its result [4]. In contrast, Freudian theory [B] explains guilt as the interaction of the id, ego and superego. The id, present from birth, conflicts with society by trying to satisfy the infant's basic needs. The ego then develops as a mediator to help satisfy these needs. Finally, the superego emerges as internal representative of social values. It inhibits sexual and aggressive impulses of the id while urging the ego to pursue socially approved goals.



Adolescence

The beginning of adolescence, the period between the ages of 12 and 20, is marked by physical changes brought about by the sex hormones [1]. These physical changes are accompanied by psychological changes. Before adolescence, most children are primarily interested in their own sex. Boys form gangs with boys, girls whisper and giggle with girls. As adolescence advances, each sex becomes more interested in, and more tolerant of, the other. But both boys and girls tend to continue to have attachments to older members of their own sex whom they admire. "Crushes" upon teachers, or the hero-worship of sportsmen or pop stars [Key] also serve a valuable function, for such people act as models that growing boys and girls can imitate, thus learning how to become fully adult members of their own sex [4]. In this way, older people from outside the family can serve as substitutes for the parental models that governed early childhood.

Physical changes in early adolescence

Early adolescence is a time when both sexes may have anxieties about the changes taking place in the body. Girls are sometimes self-conscious about their figures or bothered by menstrual irregularities [2]. Boys often worry about their developing genitals, believing themselves less well endowed than their contemporaries. Because both boys and girls develop at different rates and different ages, those who are late developers easily become anxious that there is "something wrong" with them. This anxiety is often masked by hypochondriacal fancies and many adolescents seek reassurance that they have not got tuberculosis, cancer or another complaint.

In both sexes one unpleasant side effect of the influx of the sex hormones is the development of acne. This unsightly skin complaint which can affect the chest, back and face increases self-consciousness especially if it appears on the face. Sufferers not only feel, quite rightly, that contemporaries will find them less attractive, but also sometimes believe that acne is a sign of, or punishment for, their early sexual experimentation. Acne is one adolescent complaint that requires expert medical attention.

Another is the obesity that often afflicts both girls and boys at this period of their lives. Obesity should never be regarded as just "puppy fat" that will be outgrown. For one thing, it often becomes permanent. For another, it is sometimes the result of compulsive overeating, a sympton of anxiety and depression. Like acne, this is a physical disability that is often neglected and it is important that it should be corrected.

Social awkwardness and exhibitionism

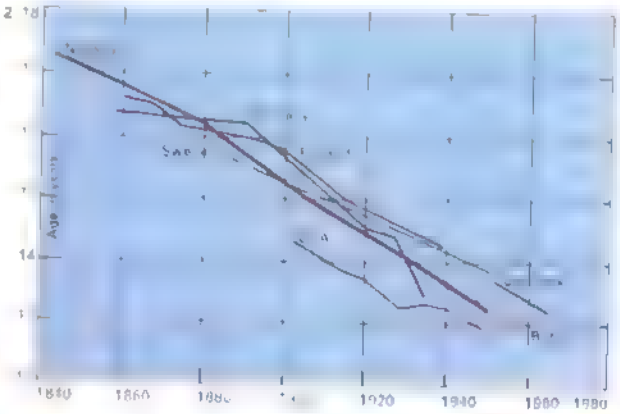
Adolescence is in any case a time of social awkwardness. To be neither child nor grown-up is a difficult situation. Some adolescents become intensely shy, hiding themselves away from social contacts. Others become aggressively exhibitionist, flaunting the start of emancipation with outrageous clothes and outlandish hairstyles [5].

Now that adolescents in Western cultures have more money they have become a mass market and the target of advertisers. Many advertisements exploit the adolescent's natural anxiety about physical appearance and social acceptability. The result is that many adolescents overspend on clothes,

CONNECTIONS

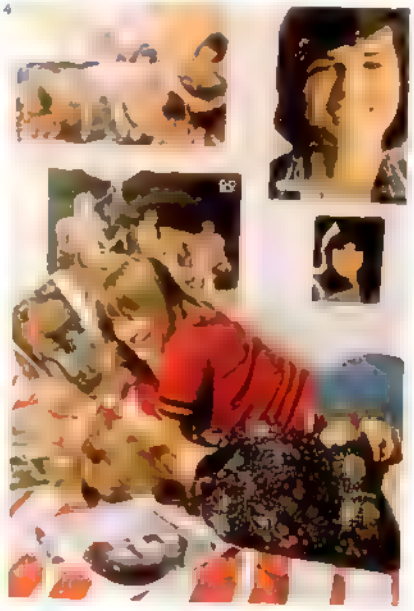
See also

1 Levels of sex hormones in the blood before birth affect gender development. Lack of balance in the sex hormone level may result in a person being in an intermediate state between male [2] and female [6]. Such a person may come up against a social block [1] and have to choose one role. Later in life intermediates [3, 4, 5] may return to emotional states and life-styles more appropriate to their own nature.



2 Over the last century menstruation has begun at an increasingly early age. Girls should be prepared for its onset by being given a straightforward account of its function and purpose. If this is properly done, menstruation can come to be looked forward to as a sign of being grown-up rather than as a "curse" to be concealed and moaned over. Usually most girls experience little discomfort.

3 Adolescence is a period when the individual tries to develop a sense of his own identity and define his position in the world. As a result, parental values [1], are often rejected and acceptance by peer groups becomes of great importance. Through increasing experimentation with the opposite sex [2] and strong attachments to boys of his own age [3], the adolescent widens his social world and learns social skills. His dress and attitudes often change to conform more closely with the group [4]. By participating in social activities such as games [5], talking, drinking and smoking [6] and parties [7] the adolescent learns to adopt different roles in different social situations. The individual passes through this learning process which helps and to his maturity before joining the adult community [8] and accepting the adult world and its responsibilities.



4 Adolescents want to leave childhood behind and become "grown-up". It is natural for them to find heroes and heroines on whom they can model themselves and about whom they can have fantasies. Society does not take enough trouble to make use of the idealism of adolescence by providing models of the right kind. Very often the adolescent heroes are themselves hardly more than teenagers and may lend a spurious glamour to such activities as drug using.

make-up, deodorants and hairdressers feeling that they must match social expectations that have been artificially implanted in their minds by the media.

Search for identity in the young adult

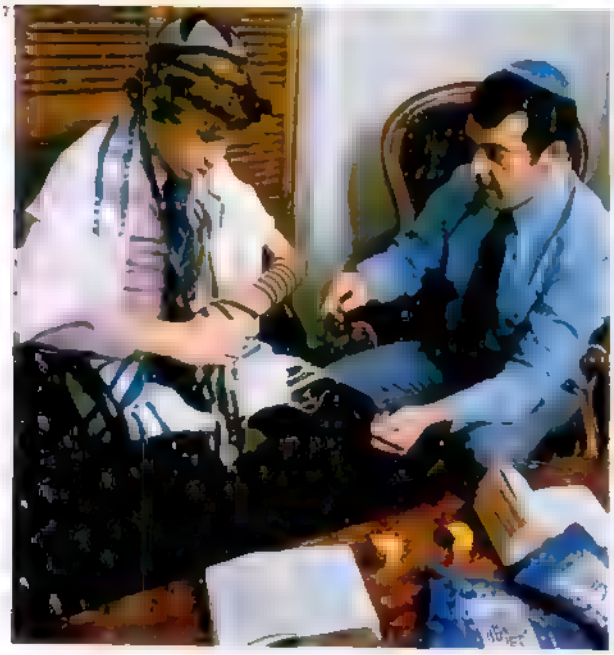
Adolescence is essentially a time of search for identity [5]. Uneasily poised between childhood and adulthood, the adolescent finds its centre of gravity in the question "Who am I?" In primitive societies, the transition from being a child to being grown-up is generally clearly marked by initiation rites, often of a painful kind. But when these ordeals are over, the individual at least knows where he is and exactly what is expected of him. In Western societies, although confirmation ceremonies such as bar mitzvah [7] may serve something of the same function, the adolescent has no clear picture of his role in society. This is partly because while some adolescents leave school at 16 and have to go out to work, others stay on the more "childish" position of being a student until well into their twenties.

In complex societies, it is much more difficult to define [illegible] since different things are expected of different levels of intelligence and background. At what age should adolescents be allowed to vote, drive cars and motor cycles, to marry, to have bank accounts? Different countries have different rules [8] and within the same country an adolescent may be put in charge of a potentially lethal machine on the roads, yet be allowed no political [illegible] in the affairs of his country or home town. Deprived adolescents who feel themselves to be undervalued or unwanted turn to hooliganism, a problem in all "advanced" cultures. If they cannot make themselves felt in constructive ways, they may consider violence an acceptable alternative [6].

Aggressive self-assertion is an inescapable part of adolescence for most of us. For how can an adolescent to define himself unless he has some standards against which to rebel, some way of demonstrating that he is an individual in his own right, a different person from his parents. Discovering identity necessarily implies discovering differences — and differences between the adolescent and the parents are [illegible]

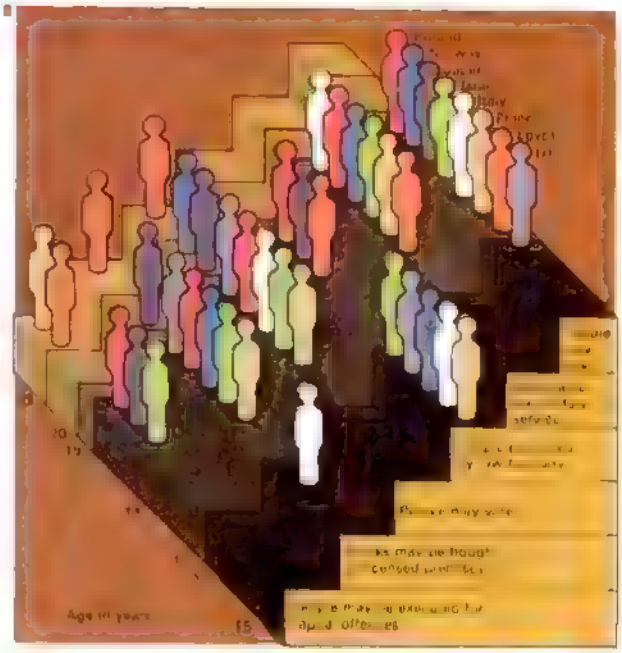
5 Clothes are one of the most obvious ways of expressing individuality, or rebelling against parents' standards and demonstrating that the adolescent is now ready for sexual encounters. Hence many adolescents dress in ways that are exaggerated and that emphasize the sexuality. Girls use elaborate make-up and boys wear jeans that emphasize the genital bulge.

6 Rebellion against parents, teachers and other authorities is a necessary part of growing up. But often it goes too far. Vandalism, gang warfare and the affrays that occur at football matches are predominantly teenage affairs. Western society is conspicuously inefficient at providing enough constructive outlets for natural adolescent energy and physical aggression.



7 Bar mitzvah, one example of a rite marking the transition from child to adult is reached by a Jewish boy of thirteen. All the commandments such as the binding of phylacteries, shown here, symbolize the Covenant, become incumbent on the boy. Confirmation and initiation rites, in simple societies and boarding schools are other examples.

8 The age at which adolescents can take on adult privileges or burdens varies from country to country as the chart shows. Teenagers are often confused because they are expected to behave like children in one setting and adults in another. Moreover the expectations are quite different in different social and intellectual strata.



Adolescence: sex and independence

Sex and independence are indissolubly linked. Sex is the biological force that compels an adolescent to look for a mate and the incest taboo ensures that this mate comes from outside his or her own family circle. The taboo on incest, which operates in almost every culture, has more than one kind of significance. First, it removes a dangerous source of disharmony within the family by excluding the possibility of sexual rivalry. Second, it links different families and thus encourages social cohesion. Third, it spurs the adolescent to become independent by forcing him or her to look for sexual experience outside the family circle. Those adolescents who, for one reason or another, stay at home may remain dependent and immature. Lying behind these social and psychological considerations is the biological possibility that incestuous unions may concentrate genetic defects in the family and produce serious mental or physical abnormalities.

Sexual experimentation

The sexual experimentation of adolescents, which often causes parents such anxiety, is in fact a biological necessity that is as much a stage in the development of the adolescent's independence as it is a means of discovery in the sexual field itself [3].

During adolescence boys and girls may experiment with both solitary and mutual masturbation, and the latter may be with their own sex or the opposite one. Transient homosexual encounters are common and should not be taken as meaning that the adolescent is permanently orientated in a homosexual direction. Nor, in most instances, should such encounters be thought to have any lasting effect. Experts believe that those who become permanently homosexual do so partly because of genetic predisposition and partly because of emotional influences during early childhood.

The changes of partner so characteristic of adolescence are also biologically and psychologically healthy. How are adolescent boys or girls to learn what their real preferences and requirements are if they are given no choice and no opportunity at all to experiment with different partners? Parents may be content if their son 'goes steady' with, and then marries, the girl next door, but psychiatrists who have seen many such marriages break up in middle age are more confident of the future security of a marriage when both boy and girl have had more experience with different partners.

The need for privacy

The essential privateness of sex is also valuable in encouraging independence in the adolescent. Parents are often over-anxious because their adolescent children do not confide in them and some go as far as to read their children's diaries and letters. Other parents become hurt and try to force their children to tell them everything, as they used to do when they were much younger. But secrets are a necessary part of growing up and should be respected as part of the drive towards independence.

Most adolescent secrets are in any case harmless and those that are not are often best dealt with by adults other than the adolescent's parents. Privacy is the right of every human being, whatever his or her age; forced confidences lead only to resentment. On the

CONNECTIONS

See also

1 Parents complain when their children are small that they see too much of them. When the children become teenagers, their parents complain that they hardly see them at all. It is of course perfectly natural for teenagers to spend most of their time with their contemporaries. There is bound to be a divergence of interest between parents and teenagers, leading the teenagers to form a sub-culture of their own. The group bond comes partly from shared interest and partly from a need for reassurance. Adolescents are notoriously unsure of themselves and it is comforting for them to find others as uncertain as they are.

2 James Dean (1931-55) epitomized the often bitter confrontation between the adolescent and adult worlds in his films *East of Eden*, *Rebel without a Cause* and *Giant*. He revealed the mental turmoil of a youth attempting to come to terms with himself and the conventions of the adult world, and his inability to communicate this inner struggle. As the champion of alienated youth, he highlighted in his roles the gap between the generations and reminded his audience of the problems of self doubt, frustration and resentment that youth suffers on the way to adulthood. In his own life he exemplified the common gulf that exists between parents and their children at this stage in life.

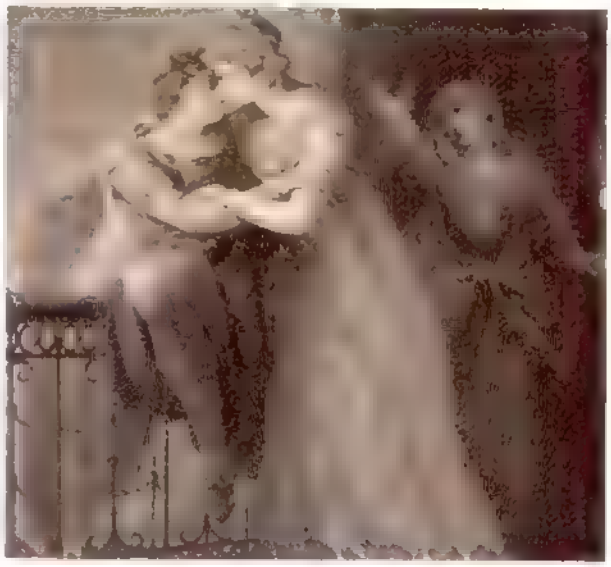
3 "The Kiss" by Gustav Klimt (1862-1918) typifies the romantic aspects of sexual involvement. Sexual potency in the male is at its height during adolescence and the urge to find sexual release is a crude insistent force that often expresses itself in coarse and vulgar ways. However these rather crude passions do not preclude more complete emotional involvement, which is commonly expressed by concern, tenderness and an unselfish preoccupation with the needs and wishes of the beloved. "Commercial sex" (such as prostitution and pornography) does not afford any opportunity for the development of emotional involvement. Adolescent interest in pornography is common, but it does not affect the normal capacity to form deep loving relationships.

other hand, adolescents also need to be understood and so may feel the need to confide in some adult - though preferably not in their parents. It is usual, if a little disconcerting, for parents to find the adolescent children of their friends and acquaintances much easier to get on with and much more polite than their own teenage children.

Adolescents typically find themselves torn by conflicting impulses. Their wish to be private vies with their need to be understood. Their wish to run their own lives is opposed to their need to ask advice about many aspects of life that they have not yet encountered. In practical terms it is difficult for them to be independent if they have to ask adults for money. And managing their own finances is a problem if they are not familiar with the intricacies of budgeting and opening an account at a bank or post office. It is difficult to maintain secrecy about a social engagement or "date" if advice and reassurances about what to wear and how to behave are desperately needed. Dependence upon parents imposes unwelcome restrictions and often results in frustration, but independence involves equally unwelcome anxiety. Only tolerance on both sides, from parents and adolescents, sees people through this difficult stage without incessant conflict [1, 2].

In place of parents

A possible solution to this inevitable and necessary tension between parents and adolescent children is the organization of society and family life in such a way that adolescents always have some known and trusted adults outside the family to whom they can turn at any time for support and advice. As an adult it is much easier to be tolerant and understanding towards the adolescent who is rebelling against parents when one is not directly involved. In the extended families of some cultures the adolescent will often be friendly with a number of adults who are concerned about him but who are not his parents. This probably makes life much easier for both parties. In the days when godparents really counted they could play the part of adviser and confessor, and there is a case for resuscitating the position of the godparent.



The story of Romeo and Juliet, members of two warring families in Verona, who defied the conventions of their society, expresses the importance of love as a force for independence.

4 Romantic and cynical attitudes to love have been conflicting elements in society since time immemorial. The romantic view is expressed in this nineteenth century Indian picture of a girl yearning for the return of her lover while Mozart's opera *Cosi fan tutte* is a tale of mutual deceit. Two girls whose lovers pretend to go into the army take substitutes who turn out ironically, to be their own lovers in disguise.

5 Opposition to the Vietnam War became a focus for young people's dissent in the 1960s and 1970s not only in the USA but all over the world. Perhaps prompted by the USA's apparently imperialistic and aggressive role in the war, there was evasion of compulsory military service on an unprecedented scale and widespread peace demonstrations that were joined by returned veterans, as here.

6 Greater affluence and a longer period of education have contributed to the emergence of an alternative mass youth culture. The wide use of illegal drugs by Western youth typifies both its affluence, independence and defiance of adult authority.

7 Clothes are a sure pointer to the profound social change that has occurred in this century. In the early 1900s adolescents seem to have emulated their parents' mode of dress. Now the reverse is true.

Adulthood: from 20 to 30

The decade from 20 to 30 is hard to characterize because there are marked differences between the roles of men and women, between the married and the unmarried and between those with children and those without. Ideally, it is a time of hope and increasing confidence. With the tribulations of adolescence behind them, and many exciting possibilities ahead, young men and women start the road towards maturity impatient to prove themselves and certain that they will do better than their seniors.

The bonds of matrimony

Yet for the newly married with a young and growing family financial difficulties are common and, particularly for women, the adjustment to a life centred round the home is difficult to make. Whereas a husband can escape to his work, his wife is often tied all day to the house. For many mothers this leads to a drastic loss of confidence and of self-esteem. Current research indicates that one-third of working-class women with children under six who are at home full time are clinically depressed.

Some women especially those whose parents have kept too tight a rein on their daughters' independence, find the sudden release from control and the new responsibilities of running a house and budgeting difficult to cope with. Some even develop phobias about going out alone and shopping and remain confined to the house because they cannot face the burdens of adulthood.

Both husband and wife may find themselves torn between the attention their partner expects and demands and their own wish to remain loyal to the group of friends with whom they spent time before getting married. It takes time for couples to form new patterns of relationships in society as a pair, and as a result some newlyweds, feel isolated and cut off from family and friends.

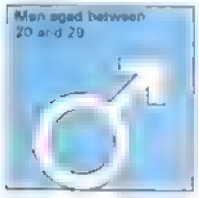
For many newlyweds money becomes an important issue for several reasons. During adolescence young people, at least in Western society, are often still living in their parents' home and yet have begun to earn their livings. Before marriage, they have few financial responsibilities and become accustomed to spending what they earn on themselves without thought for others. After marriage, their financial horizons become sharply restricted. The couple may soon acquire a home of their own, but they may have the prospect of many years of regular repayments on the house or apartment and on the furniture to fill it. A home means privacy and comfort, but it also means a succession of bills to be paid.

The arrival of children

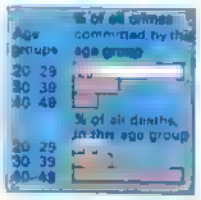
The situation may become more complicated after the arrival of children because they add to expenditure and because after the birth of the first child many women give up their jobs. This reduces the family's income and at the same time promotes a feeling of dependence in the wife. The fact that the money she spends is earned by her husband may seem to her to be irksome and degrading, particularly as she makes such a contribution to the quality of the home life – a contribution that some women think is taken for granted or undervalued. This can become a focus for resentment and bickering. Equally, some men feel that the responsibility of being the

CONNECTIONS

See also



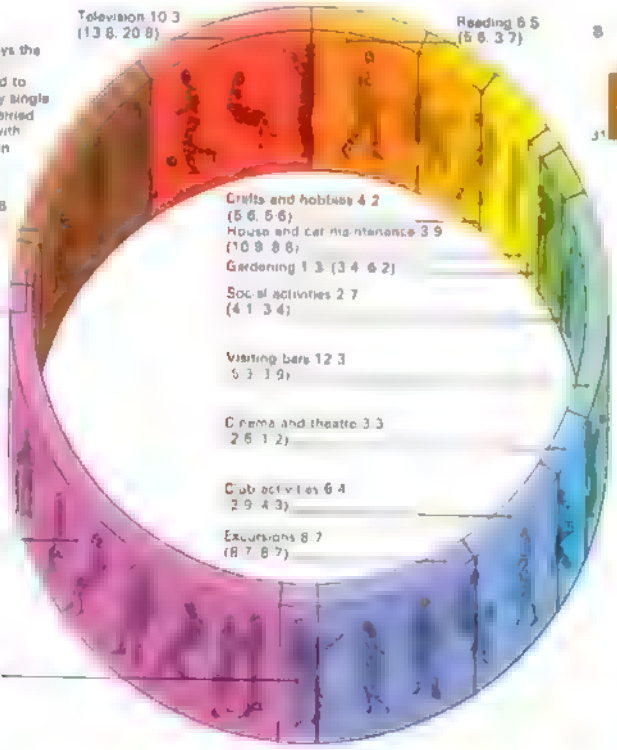
1 The average single man [A] in Western European countries spends a quarter of his leisure time participating in physical recreation. The most popular pastimes are swimming, football, table tennis, athletics, tennis and rugby football. Once married he adopts more sedentary pursuits centered on the home. The arrival of children further consolidates the family unit. Salaried workers [B], who are often under training in their early twenties, on average earn less than their manual worker counterparts although they have greater earning prospects in the future. Men between the ages of 20 and 30 committed almost a quarter of all crimes in 1974 [C]. However male teenagers were guilty of 26 per cent of all offences, among which car and petty theft were most common. The greatest killer [D] of 20 to 30 year olds is the motor car, causing 30 per cent of all deaths, while suicides account for 10 per cent. Cancer is the major medical cause (leukaemia and cancer of the testes particularly), heart disease ranks second.



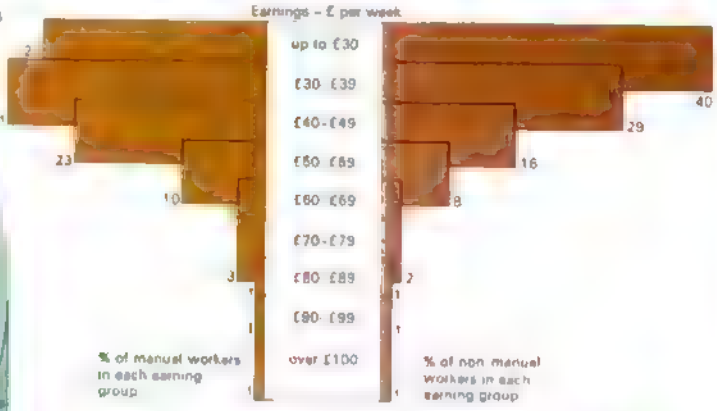
1 Leisure activities

A

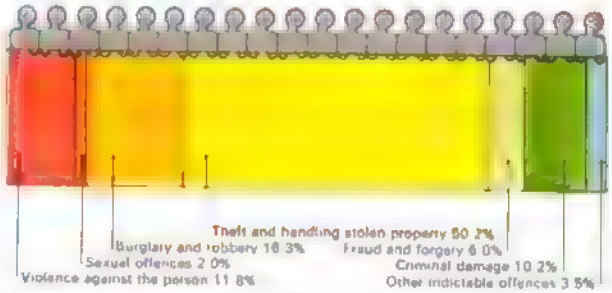
The diagram portrays the percentage of total leisure time devoted to various activities by single men. Figures for married men without and with children are given in brackets.



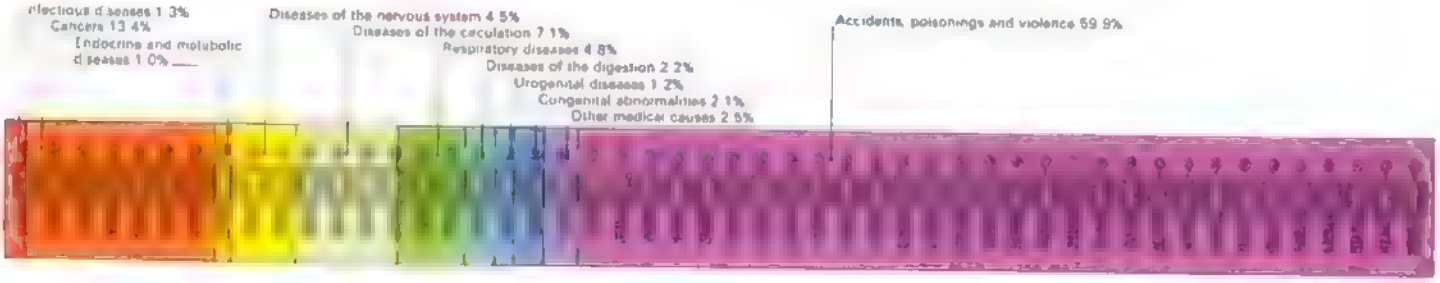
B



C Crime – indictable offences



D Causes of death



sole supporter of the household is a heavy burden. The classic role of the man as breadwinner can be as constraining and unwelcome as the role of the woman housewife.

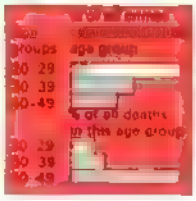
In addition children impose financial strains and there is evidence that they may also precipitate problems of communication, particularly for low income families. While husband and wife both have jobs they share many preoccupations. But as the wife becomes absorbed in her work in the home and the husband in his work outside, these common experiences largely disappear and there is often little to talk about of mutual interest. For many couples this is an important factor in marital disharmony.

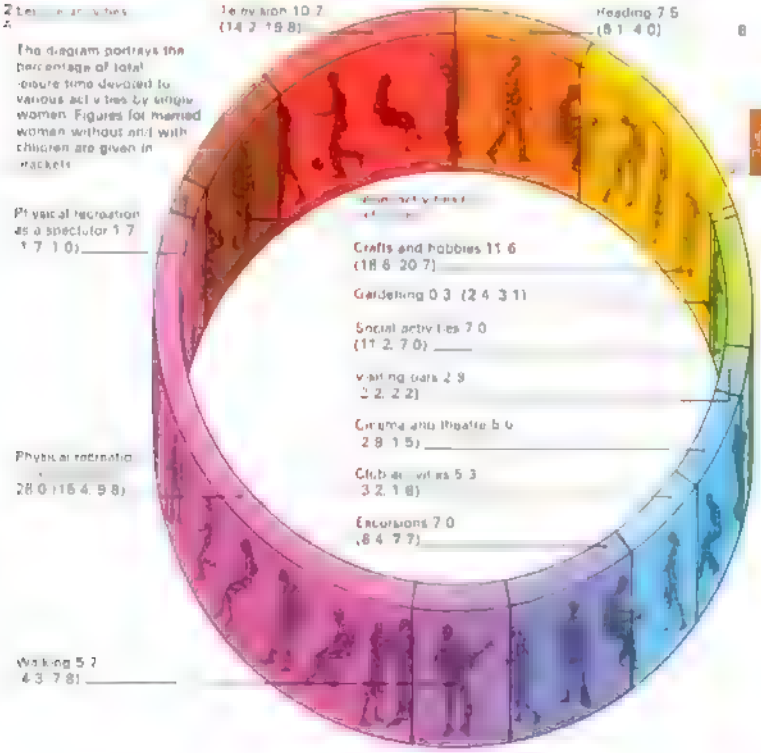
The constraints of the period

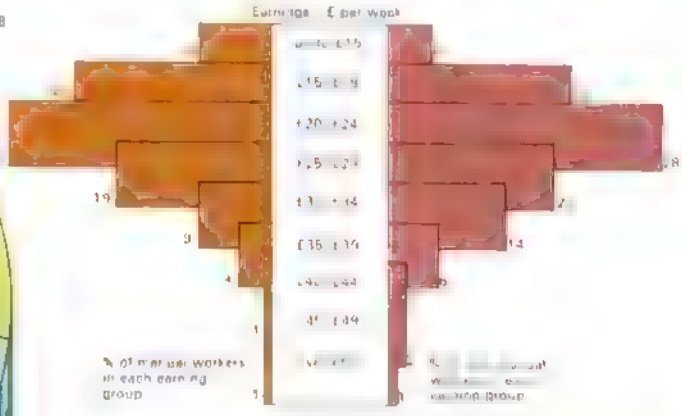
Western societies do not make a clear enough demarcation between being a child and being a grown-up, with the result that the adolescent often does not know what society expects of him. This problem continues in the decade 20 to 30. A particular difficulty is that although at the peak of their powers young men and women find difficulty in expressing their full potential. In a society that is still controlled by the middle-aged they cannot usually gain any position of power. In the medical profession, for example, a doctor aiming at specialist training may not have completed that training by the age of 30 and may not have attained a position of eminence until middle age has overtaken him. In societies where hunting and fighting are the main occupations of younger men, maximum potential and maximum achievement coincide. In Western society youth often lacks the opportunity to use its ability.

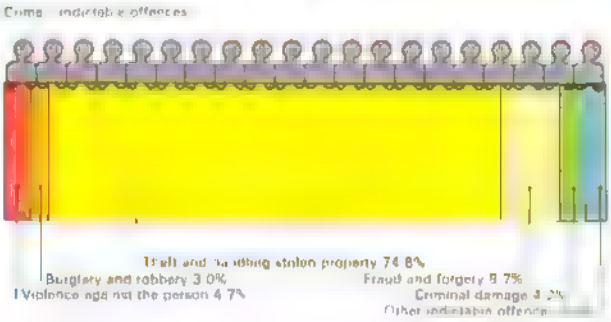
This dilemma may be particularly acute for women. The pressure that society exerts on them to measure their success in terms of the children they produce is at variance with their growing wish to participate in the world on an equal footing with men. Their traditional goals clash uncomfortably with their new ones. As a symptom of this there is a shortage of job opportunities and of such facilities as day-care centres for pre-school children suited to the needs of women who wish to combine having a job outside the home with caring for their children.

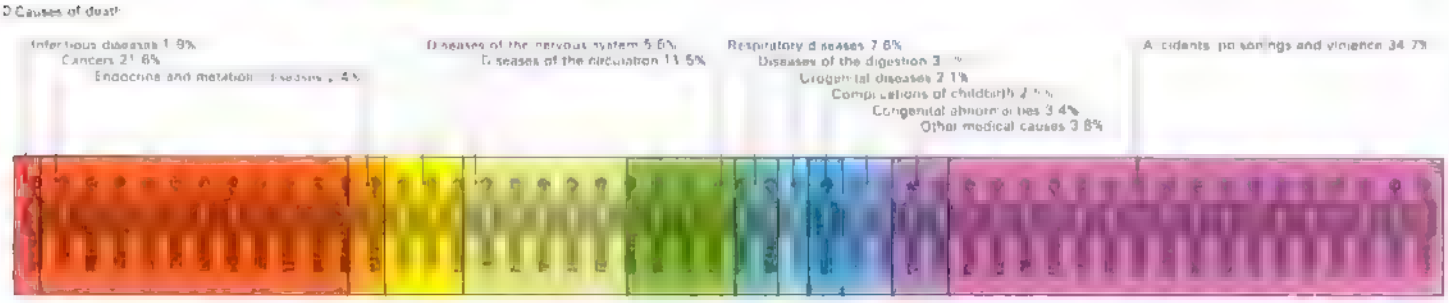
2 The most popular forms of physical recreation (A) for women between 20 and 30 are dancing, swimming, tennis, table tennis, horse riding and ice skating. The amount of leisure time spent on these activities declines with marriage and the arrival of children. As with men, activities then centre on the home and hobbies and craftwork such as knitting and sewing. The earning potential (B) of manual and non-manual women workers varies little as few non-manual workers undergo long training. There is still a wide difference between the earnings of men and women. Women aged between 20 and 30 (C) commit less than a fifth of the number of crimes committed by men of the same age group (girl teenagers have a higher crime incidence). Shoplifting comprises 50 per cent of offences but violence is increasing. The death rate (D) for women of this age is less than half that for men and deaths due to accidents are considerably lower. Relatively more women die of leukaemia, brain cancer and certain heart diseases.











Adulthood: from 30 to 40

During their thirties most young people become homes, establish careers, and adult roles. The direction which their lives are going and the probable range of his achievement will be settled, although surprising changes of direction can still occur. Equally, a woman will usually have decided on the balance she would like to strike between career and family. A married couple will know how many children they want and their social life will have been established, although this is largely dependent on the age of their children. Many older people look back on this decade as a time of particular happiness. The uncertainties of youth are past, the problems and upheavals of middle age are still in the future. Asked what age they would like to be, many older people answer: "Thirty-five."

Friendship and love

Although friends can be made at any age, by 15 to 40 most people will have a number of established friends, people whom they know so well that they can be completely at ease and relaxed with them. This is partly the result of having overcome the social anxieties of earlier decades and partly because many friendships do not fully mature for some years after the first encounter.

Nevertheless, this decade may present problems. Married couples usually find that the initial delight of being "in love" has worn off. Being in love is a curious condition which Sigmund Freud (1856–1939) called "the psychosis of normal people." Most young people carry within their minds a somewhat idealized picture of the opposite sex, derived in part from parents and in part from the images presented by television, the cinema, novels, advertising and so on. When they fall in love, the beloved person seems to correspond exactly with this idealized picture. Psychologists would say that the idealized image is "projected" upon the other person. Living with another person day in, day out means, however, that each becomes aware of all the ways in which the image and the reality fail to correspond. Real people are not like creatures of the imagination but are human beings with faults and foibles as well as the attractions that first gained their partner's interest.

Being in love, therefore, must be replaced by caring for one's partner as a real person. For some, this is an easy transition; for others they miss the thrill of being in love and may sometimes become disillusioned or bored with the person they once adored. This is the prime reason for what is called the "seven-year itch", the tendency, especially in men, to look for sexual partners outside the marriage. Infidelity, a prime reason for divorce, is now recognized as a symptom of marital disharmony rather than a cause, and transient infidelity ought to lead to a re-examination of the marital relationship rather than to its immediate break-up.

Psychological maturity

In Western culture, the years from 30 to 40 are commonly regarded as those in which "maturity" is reached. But this prized maturity is hard to define. Both men and women reach their physical peak in their early twenties or earlier. Yet often psychological maturity seems to be a goal that recedes the nearer a human being approaches it.

CONNECTIONS

See also

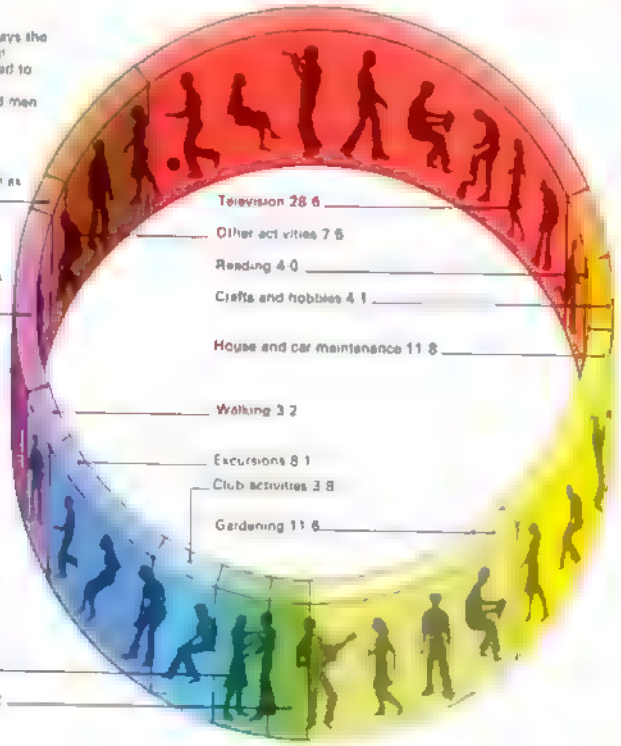
[illegible]

Adulthood: middle [illegible]

1 The average married man in Western European countries spends 60 per cent of his leisure time away from and around the home. Television, gardening, decorating the house and maintaining the car are main forms of relaxation. Time spent in physical recreation drops to only 10 per cent and favourite sports include swimming, dancing, fishing and golf. Going out in the evening is of minor importance. Earning comparisons (B) show non-manual male workers earning on average 22 per cent more than their manual counterparts, reflecting the earning potential of qualifications gained earlier. Far fewer crimes (C) are committed by men in their 30s than by those in their 20s. Destructive crimes give way to more offences of other types but only some kinds of sexual crimes rise absolutely. Male deaths (D) in the 30–40 age group are virtually the same as for younger men. There is a drop in deaths as a result of accidents. Heart disease and cancer of the lung, brain or testes are the major medical causes of death.

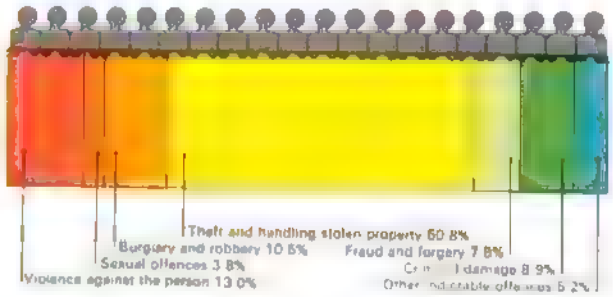
1 Leisure activities

A The diagram portrays the percentage of total leisure time devoted to various activities. Figures for married men with children.

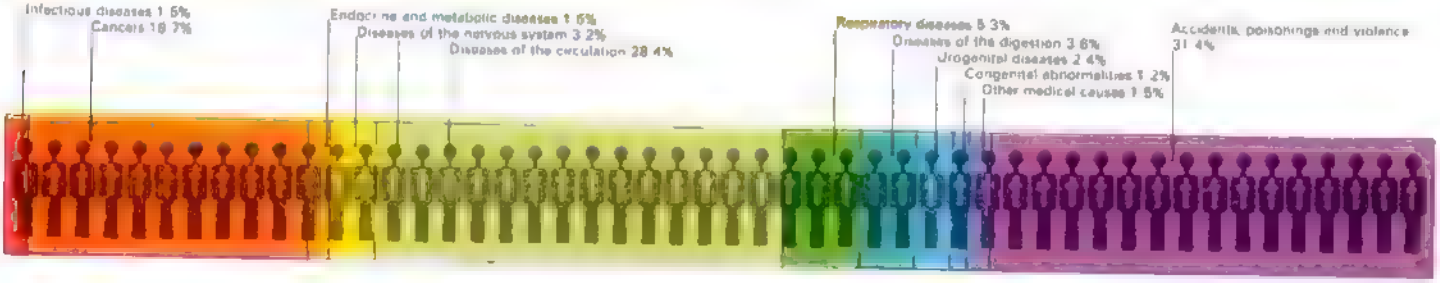


B

C Crime – indictable offences



D Causes of death



What is maturity in psychological terms? Since man is a social being, maturity of personality cannot be defined except in terms of relationships. A person without relationships with other persons cannot be defined as an entity, just as it is meaningless to ask what size an object may be if there is nothing with which to compare it. Although there are many attributes of maturity, one that is obviously important may be singled out: the ability to make fruitful loving relationships with other people on equal terms, without either being dominated or dominating. This achievement implies an acceptance of the other person as he or she is, without any wish to alter, to direct, or to submit, a recognition of the other person as a separate entity and therefore of oneself as a separate entity also.

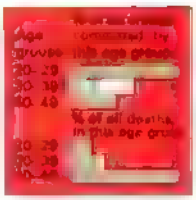
Importance of self-criticism

Maturity also demands that a person be realistic, without abandoning the capacity to use imagination. It implies self control, but combined with the ability to "let go" when this is appropriate. It implies having arrived at some fairly coherent point of view about the universe and the place of man within it, but excludes dogmatism and fanaticism. It implies firmness without rigidity, love without infatuation, decisiveness without being dictatorial, tolerance without a facile permissiveness.

In fact, to be mature in every sense is beyond the reach of most, though this ideal state may be approached more nearly by the habit of vigilant self-criticism. Few human beings in modern cultures become mature by the age of 40. Indeed the extension of the human life-span beyond the reproductive period has necessitated the extension of psychological development beyond middle age well into the second half of life.

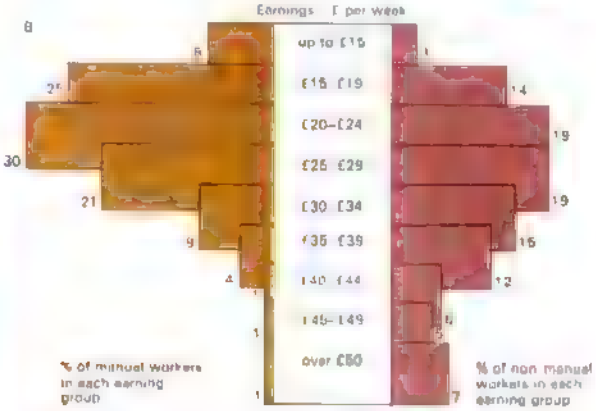
Statistics [1, 2] show that people in the 30 to 40 age group are usually healthy, socially stable and tend to spend much of their spare time within the family circle. The material rewards of having gained higher educational qualifications become more evident. In both physical and psychological terms, adults of this age are most likely to store up future trouble by developing a false sense of complacency.

2 Married women in the 30 to 40 age group [A] spend little time on physical recreation apart from dancing. As with men, more than 60 per cent of leisure time is spent in the home and going out is confined mainly to places the children appreciate. Television, reading, knitting, sewing, other handicrafts and gardening are the major activities. Earnings of women in non-manual work [B] rise in the 30 to 40 age group and average 33 per cent more than for women in manual work whose earnings are no higher than for the 20 to 30 manual group. Women's crime rates fall in the 30s [C] compared with the 20s. Petty thieving accounts for 80 per cent of offences, of which shop lifting, the only offence committed more by women than men, is 55 per cent of the total. The death rate of women in their 30s rises faster [D] than it does among men. Cancer of the breast, ovary or cervix, together with cerebrovascular and rheumatic heart disease, are by far the main causes of death in their group.

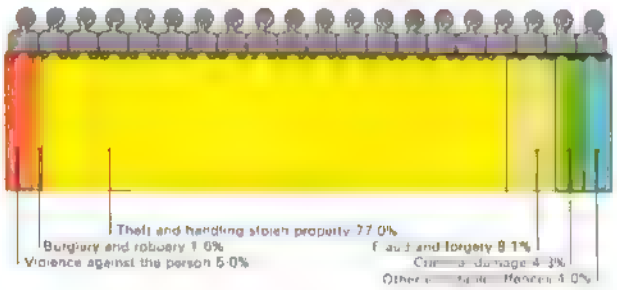


A Leisure activities

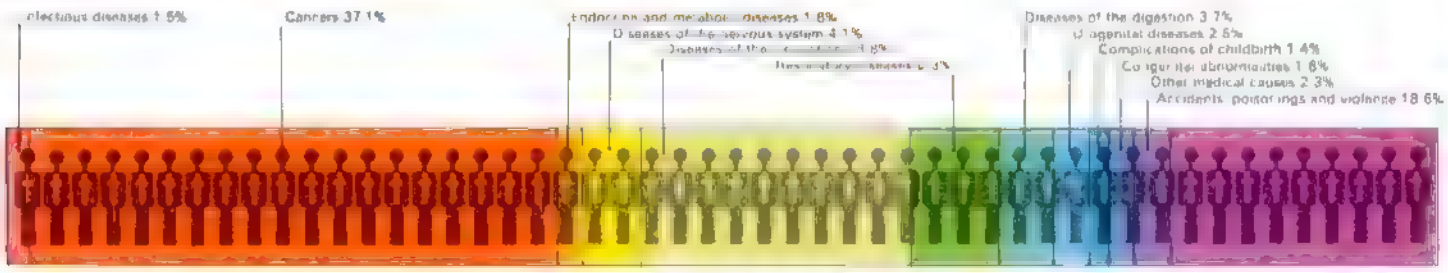
The diagram portrays the percentage of total leisure time devoted to various activities. Figures for married women with children.



C Criminal indictable offences



D Causes of death



Adulthood: middle age

In modern culture there is a growing awareness of middle age as a period of change and development. It was once assumed that by the age of 40 both men and women were so fixed in their ways that little change could be expected and to an extent this is borne out in statistics [1, 2]. But it is increasingly recognized that middle age can and should be a time in which new interests are developed and new departures undertaken.

Mid-life crisis

Emotional disturbances in middle age are so common that psychiatrists often refer to the "mid-life crisis". This was once thought to be the result, at any rate in women, of the "change of life" or menopause. While it remains true that a few women do experience unpleasant physical symptoms at this time and may become irritable and moody, modern treatment with hormones, although controversial, is able to alleviate most of the symptoms. Moreover, the widely held belief that after the menopause women lose most of their sexual desire is demonstrably false. In fact the opposite is often true, since many women experience an increase in desire when pregnancy is no longer a possibility.

Although men often go through a mid-life crisis, this is no longer generally believed to be the result of some male equivalent of the menopause. It has been shown that diminution in sexual desire and performance is very gradual in most men, so that the emotional upheaval through which some pass cannot be attributed to any sudden decline in sexual potency. What then is the cause of the mid-life crisis, and what forms does it take?

The name itself suggests one reason. At the mid-point of life, both men and women are at the stage when they should have realized their ambitions. If things have gone well, the traditional goals (for a man, of establishing himself in a job or profession, for a woman, of rearing a family) will have been achieved. However, many women feel there is nothing left for them once their children are no longer dependent and that years of child-care and housework have left them unqualified for work outside the home. And many men regard the effort they have put into "getting ahead" as misplaced.

Inevitably, not all of a person's dreams will have been fulfilled. Some become depressed because they have to come to terms with the reality of what their life is, and abandon some of the hopes of youth. This is a time when a restless dissatisfaction afflicts many people, often showing itself in transient infidelities, increasing consumption of alcohol and changes of occupation. People of this age often feel that the future has nothing to offer but decline into old age and death.

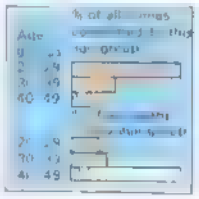
Creative turning-point

A study of the lives of creatively gifted men and women shows that their work often undergoes a change of style at the mid-life period. Some, like the English novelist George Eliot (1819–80), do not begin their real creative production until this time in their lives. Others, like Ludwig van Beethoven (1770–1827), demonstrated an increase in profundity that marked off their late work sharply from what had gone before. Sigmund Freud (1856–1939) published nothing of striking originality until he was 39, and the work that he continued to regard as

CONNECTIONS

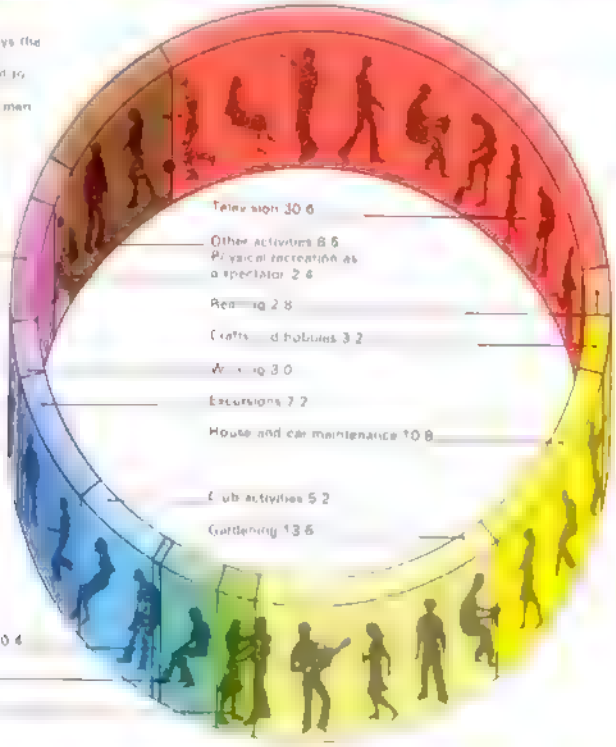
See also

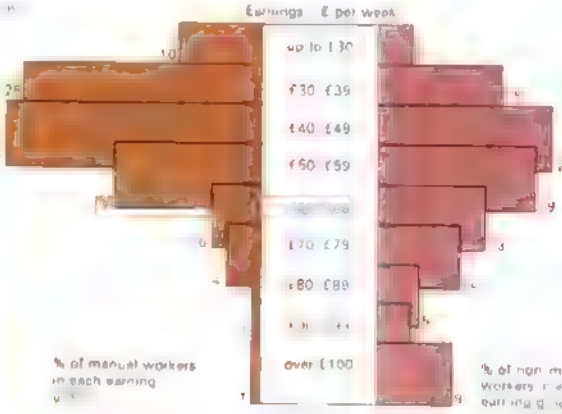
1 Leisure activities
[A] of average middle aged fathers in Western European countries continue to centre around the home, although children are likely to be more independent. Gardening now takes precedence over the house. Less energetic sports are taken up and watching sport is very popular and often shared with sons. Earnings of male manual workers [B] show a drop in middle age because of declining productivity but the non manual worker increases his earnings and his relative prosperity. Crime rates [C] for men aged 40 to 50 show that they commit fewer impetuous offences than younger men. Most of the crimes in this age group are committed by professional thieves. Men of this age group are responsible for only 4.8 per cent of crimes. Typical causes of death [D] in an affluent society – heart disease and cancers – account for more than two thirds of all deaths of middle-aged men. Mortality rates in general begin showing a rise between the ages of 40 and 50.



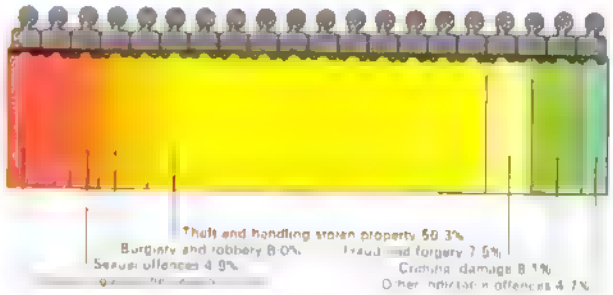
1 Leisure activities

The diagram portrays the percentage of total [illegible] devoted to various activities. Figures for married men with children.

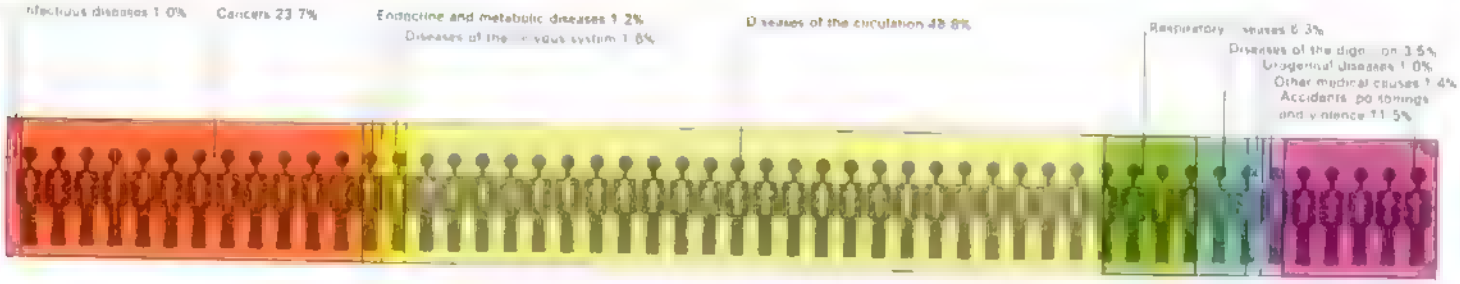




C Crime – indictable offences



Causes of death



his most penetrating *The Interpretation of Dreams*, did not appear until he was 43.

Creative geniuses show in their work that they suffer from the same conflicts as other people. Their ability to cope with this, sometimes by changing their style, is something that others might emulate.

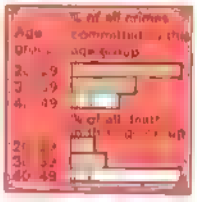
Discovering new goals

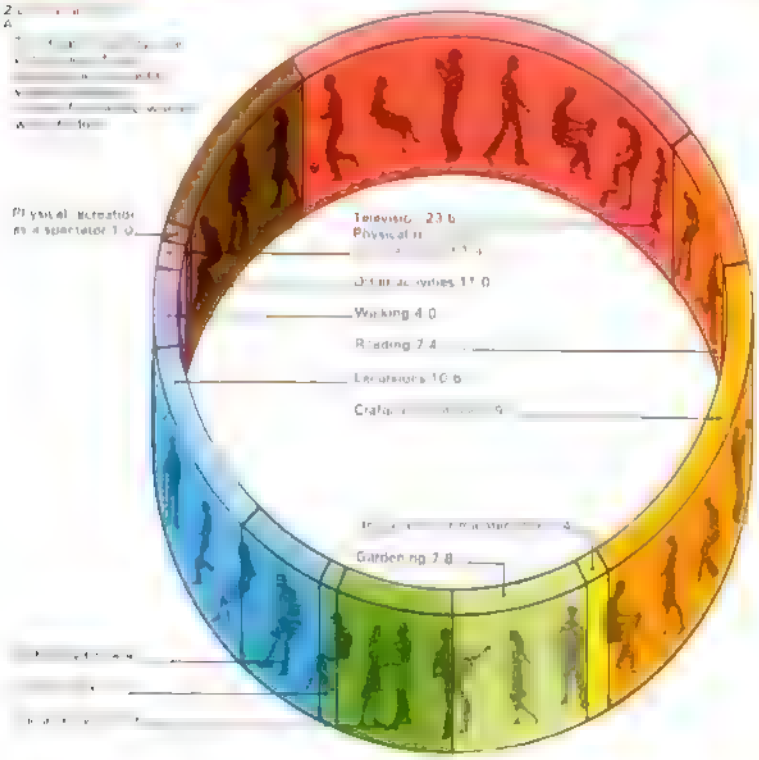
Often the mid-point of life is a time to rediscover interests and aspects of the self that have been dropped because there has not been enough time to pursue them. Most middle-aged people have had enthusiasms in adolescence that they were unable to follow up – for painting, music, literature, gardening, bee-keeping or bird-watching. The great Swiss psychologist Carl Jung (1875–1961), who specialized in the treatment of middle-aged patients, said that culture was the goal of the second half of life, and that what we needed were schools for 40 year-olds. It was a perceptive and perhaps prophetic suggestion. Middle-aged people sometimes fear that they cannot learn anything new, but this is wrong. At this time most people are more realistically aware of their strengths and their weaknesses, and are much more able to apply themselves systematically to whatever they undertake.

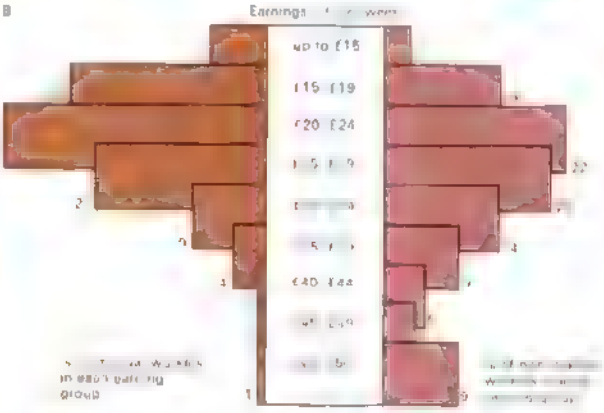
The great mistake is to think that in middle age it is too late to continue searching for anything new. For a while, psychologists thought that the human being was always driven by a need to discharge tension and achieve the peace that comes when instincts are satisfied and hungers temporarily assuaged. Now it has become clear that such an idea is quite inadequate. The brain works best when it is given a variety of different, novel stimuli. Shut a man up in a darkened, sound-proof room with nothing either to worry or disturb him, and what happens? He has hallucinations and begs to be released.

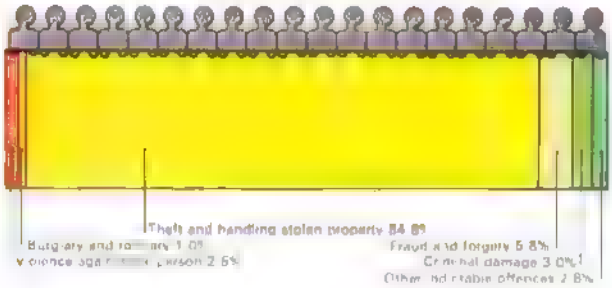
New problems are a vital part of living, and if they do not exist it is necessary to invent them. Men and women who have achieved, at the mid-point of life, the goals of youth, must find new problems to wrestle with, and other interests to engage them. Discovering these fresh challenges is largely a matter of having the right attitude.

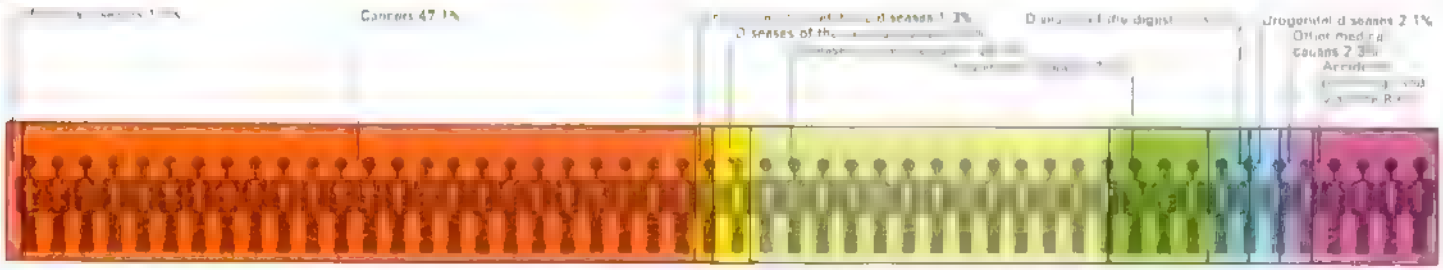
2 Leisure patterns A) among middle aged women show a slight fall in television watching and hand crafts. The extra time is spent in and around the home in gardening or in reading. With a greater degree of freedom from children there is an increase in physical recreation, especially dancing. Women engage more in social activities than men. Earnings of female manual workers (B) are maintained at much the same level in middle age as they were in the decade 30 to 40 but they are still consistently less than a man receives for the same work. Few crimes (C) are committed by middle-aged women. Shop-lifting and other petty thefts account for an overwhelming percentage of their offences. Drunken driving is the only offence more common than among younger women. The death rate (D) among middle aged women climbs steeply. Cancers and heart disease cause more than 70 per cent of deaths at this age. Of these breast and cervical cancers predominate.











Ageing and longevity

The rate at which human beings age has not changed since prehistory. Civilized man lives longer [2, 4] because his chance of survival to the age of natural death has improved, not because ageing itself has been postponed. Ageing is the term for the process that causes organisms to become more liable to die the longer they live. With time, the capacity for self-repair and resistance to damage decreases and the force of mortality increases [Key] at a rate characteristic of each species, leading to a characteristic life-span.

Longevity, size and inheritance

The life-spans of animals [1] tend to correlate both with their size and, more closely, with excess brain capacity (as calculated by the information capacity carried in the nervous system) in relation to size. The human life-span is often quoted as the biblical "three score years and ten", although certified ages up to 112 are documented.

In several localities there is a popular belief, recently widely publicized, that people live to great ages. The best known of these are Hunza-land, Abkhaz, in the Georgian Republic and Vilcabamba in Ecuador. Abkhazian life-spans, some of them up to 140 or 160 years and more, have been much investigated and much disputed. They can neither be accepted nor rejected out of hand. The Vilcabamban records, better authenticated than most, have been under study. There are claims of high ages elsewhere, chiefly in mountain areas where they are linked with "simple life", isolation from disease and an absence of reliable documents.

If such longevity is true, the causes are likely to be in the first place genetic, with social and possibly dietary factors playing a part. Longevity in humans is heritable, but less so than is stature. What is inherited is not so much longevity as the absence of factors for "shortevity". In most known genetic systems, hybrid vigour (the inheritance of a varied genetic repertoire) outranks any other heritable factor in increasing life-span.

Most multicellular organisms age, the probable exceptions being those, such as large trees and sea anemones, in which all of the component cells propagate clonally by splitting into two exactly similar cells. Ageing occurs in mixed organisms such as mammals, which contain both renewable and unrenewable cells, and in organisms (such as certain worms and adult insects) in which no cell division is possible.

The process of ageing

The nature of the ageing process or processes is unknown. One question to be settled is whether in man and other mammals the accumulation of damage is sited primarily in non-dividing cells such as neurons, in dividing cells, which may develop faults with successive divisions, in structural materials such as collagen (a protein in fibrous connective tissue); or in the programming of our physiology. No one of these processes can yet be pinpointed as the "cause" of ageing.

Fixed cells accumulate both viruses and waste materials and are irreplaceably lost with time, although the widely cited notion of massive brain cell loss with ageing is erroneous. Dividing cells accumulate mutations and errors but, except when they escape from normal controls, as in cancer, countervailing mechanisms operate to censor

CONNECTIONS

See also

1 The only mammals comparable in longevity to man are the elephant and the rhinoceros. Small rodents live one year in the wild but may survive two to four years in captivity. Dogs seldom reach more than 20 years, cats possibly longer. About 30 years is the maximum age for most medium or large mammals. Whales are the only possible exception, although the evidence is confused by the difficulty of interpreting the zones laid down in the wax ear plug (as in tree trunk rings): these are annual. Ages of 80 years may be reached by some species. Birds live longer than mammals. Small species, normally annual in the wild, may live 12–20 years in captivity. Large flying birds (parrots, geese, ravens) may be capable of living 60–70 years. The fish with the longest potential lives are sturgeon (100 years). Animals are listed with their known life-spans in years: [1] seahorse 1–2, [2] mouse 3, [3] guppy 5, [4] guinea pig 7+, [5] queen bee 7+, [6] silverfish 7, [7] large beetles 5–10, [8] earthworm 5–10, [9] swallow 9, [10] bats 10–15, [11] sponge 15, [12] rabbit 12, [13] sheep 15+, [14] frog 12–20, [15] starling 19, [16] giant spider 20, [17] dog 24+, [18] seal 20–25, [19] cat 27+, [20] cow 30+, [21] snakes, lizards 25–30+, [22] oyster catcher 27, [23] pigeon 35, [24] lion 30–35, [25] newt 35, [26] toad 36, [27] zebra 38+, [28] chimpanzee 39+, [29] ostrich 30–40, [30] horse 40+, [31] hippopotamus 54, [32] carp 50+, [33] lobster 50, [34] pelican 40+, [35] mussel 50–100, [36] goose 47+, [37] crocodile 50–60, [38] sea anemone 60–70, [39] Indian elephant 77+, [40] cockatoo 70–85, [41] golden eagle 80, [42] sturgeon 80–100, [43] man 112+, [44] vulture 117, [45] big tortoise 100–150

mismade cells. It was once believed, from faulty experiments of the French biologist and surgeon Alexis Carrel (1873–1944), that cells grown outside the body were "immortal", being capable of indefinite division. Recent work indicates that this is not so, but that normally they are able to undergo only some 50–60 doublings before the accumulation of defects in their chemical machinery kills them. Major changes also occur with time in the body's immune defence mechanisms. The probability here is that divergence in cell structure causes some cells not to be recognized as "self" and so they are destroyed.

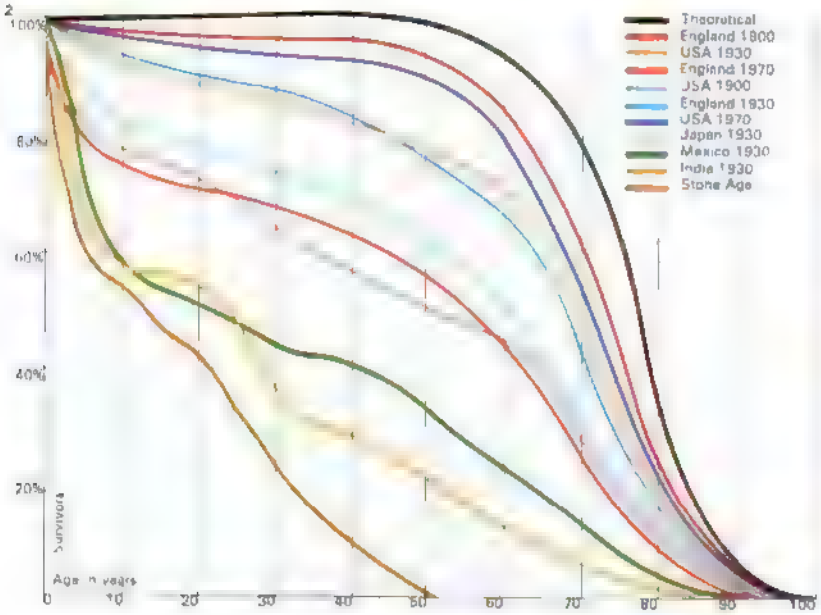
Ageing in most animals can be readily modified by the simple process of limiting calorific intake [5]. This limitation need not be so severe as to affect growth; feeding rats two days in three has little effect on growth but increases life-span by 60 per cent or more. Such observations suggest that the rate of ageing in man could be retarded.

Since mechanisms in the brain probably monitor dietary intake and serve as controllers of cellular ageing rates, the study of these, rather than of lifelong dietary modification, seems a likely course for research aimed at making it "take 70 years to reach 60". Human life can probably be prolonged little by overcoming single diseases, with the possible exception of arterial disease; a man dying at 40 usually has one cause of death, a man dying at 90 may have 15.

Towards longer life

Lifetime dietary or other experiments on man or long-lived animals could be useful but this kind of experiment is in practice confined to rats and mice. High priority in research on ageing is therefore given to tests of short duration. Such tests were first applied to Hiroshima survivors in an attempt to find whether ageing is accelerated by radiation. Some animal studies designed to discover whether anti-radiation drugs delay ageing have prolonged life in mice, but it is likely that changes in food intake or liver chemistry have played a part. A technique to alter the human ageing rate is however almost certainly attainable [3], and much research is being done under the title of gerontology.

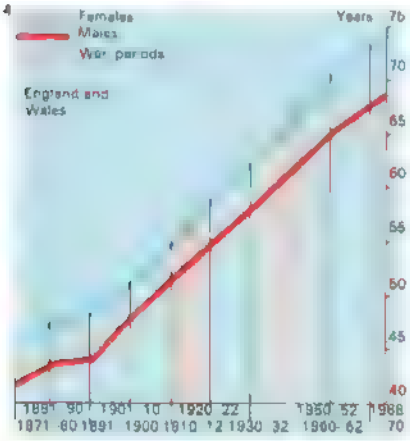
"You start dying the moment you are born" is a favourite saying caught in this artist's impression of man assailed by death. If we dodge the rocks, arrows, blunderbuss or rifle of death, the bridge itself stops halfway across the river.



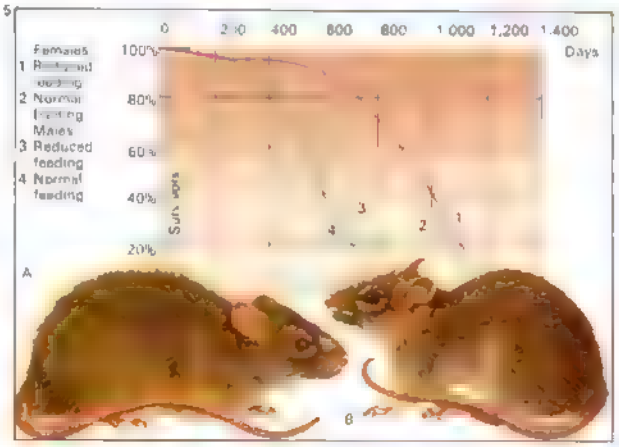
2 Human survival curves show the effect of improvements in living conditions on the lengths of time that people could expect to live in different cultures and at different times. The shape of the curve is revealing. A rapid plunge at the start, as in Stone Age man (an estimate), Mexico (1930) and India (1930) indicates high infant mortality. The top line is the theoretical optimum, a situation in which almost everyone lives to be about 70 years old. The curves for modern industrial nations are approaching this point. Current research on ageing is aimed at achieving the optimum, where active life is prolonged to the limit. The object is not to produce a population that will live to the age of 120, but to develop one that will retain its vigour up to a natural life-span of about 80 years.

3 Many interlinked factors probably contribute to the ageing process. Certainly no single common factor has been isolated. Consequently the research that is being carried out in several countries is aimed at testing many different areas that are thought to be implicated. Listed below are some of the most important. Generally tests are carried out on such animals as mice and positive results that can be applied to humans are years away.

Agents	Effects	Examples
Antioxidants	Prevent cell damage by [illegible] free radicals. Many antioxidants affect appetite, so cut down food intake and [illegible] enzymes	BHT (butylated hydroxytoluene) [illegible] ETQ (ethoxyquine)
Radioprotectants	Prevent chemical changes that are similar to those produced by radiation	SH (sulphydryl) groups
Immunosuppressants	Prevent the body's immune system from [illegible] war on the body itself	Azathioprine (used in prevention of [illegible] rejection), ALS
Cell or lysosome stabilizers	Prevent cell damage caused by enzyme leakage within the cell itself	Prednisolone (synthetic steroid induces same effects as adrenal gland), aspirin
Enzyme inducers	Cause the liver to produce enzymes that affect the body's metabolism	DDT, phenobarbitone, BHT, Dilantin
Antiappetants	Reduce food intake and thus restrict calories	BHT
Hormones	No one hormone controls ageing, but many are able to affect [illegible] different physiological processes	
Patent medicines	Most are wholly inactive although high claims are made	Gland extracts, embryonic cells, quassia, bee jelly, etc



4 Life expectancy at birth has risen markedly in industrial countries in the last 100 years, due to better nutrition, advances in medical science, smaller families and improved health and hygiene measures. Total births of boys normally exceed those of girls by about 6 per cent. But still-birth and mortality rates at almost all ages are higher for men, so typically women live longer and outnumber men in the total population (by 106 to 100 in 1971 in the UK).



5 Overeating seems to be one cause of ageing, judging by a series of experiments carried out on rats. One group was raised on a restricted but nutritionally adequate diet, others from the same litter were fed normally. After 24 months the normally fed ones (A) had effectively reached old age after only two-thirds of their lives. The "retarded" ones (B) when re-fed were as fit at 39 months as the normal rats were at 24. The survival curves reflect this difference.

An active old age

In pre-industrial societies the old were considered to be repositories of wisdom. Because they had seen it all before, and had practical experience of how recurring situations had been effectively dealt with, they were a valued and valuable resource to the community. In societies characterized by large, extended families of mixed ages the old had vital functions, not only as guardians of experience and lore but also as those who taught children and watched livestock.

Cultural prejudice towards the aged

In many contemporary societies families are smaller, their range of activities has contracted and jobs have become more specialized. Money has taken the place of skills and general functions have been usurped by specialists. Increasingly, in societies that have made youth an obsession, the old have found that ordinary dignities and privileges as equal human beings have been denied them. In the East both life and death are considered to be part of the whole life cycle and so the older a person gets, the wiser he is thought to be [4]; in the West attitudes are different. The tendency, particularly since the 1960s, has been to downgrade the importance of experience, to question the authority of those with seniority in professional or political life and to inflate the value of youthfulness.

While physical changes in ageing, such as greying, wrinkling and muscular weakness, are undeniable [1, 2], much of the picture of "old age" in Western cultures is based on social attitudes. Old age is a role imposed by a convention that assumes "the old" to be infirm, unemployable, ineducable, asexual and dependent. All of these assumptions depend on old people playing the expected role. Too often people equate retirement from a job with retirement from life and treat others and themselves accordingly. But many of the supposed disadvantages of age have been shown to be imaginary [3] or to arise from the expectation that they will occur. Although old people are not so readily aroused, sexual function persists lifelong in normal individuals of both sexes. The rate of learning a foreign language for the first time is identical in 15- and 80-year-old subjects if those with impaired brain circulation or other overt diseases are excluded. The "dependency" of the old, like the deficiency once ascribed to black people or to women, is a product of folklore and prejudice, not a fact.

The existence of a large population of aged is really a twentieth-century problem, since until recently most people did not live past 65. In America, for example, only three per cent of the population was over 65 in 1900, today the figure is ten per cent and it may soon approach 20 per cent. This is a pattern that is evident in all societies with zero population growth and it is now precisely these societies that exclude older persons from full participation and cultivate negative attitudes to their usefulness and worth.

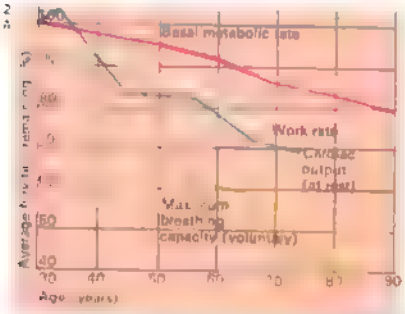
Planning for the aged

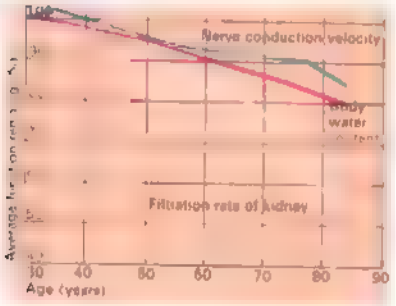
There are some signs that the body of prejudice towards older people may change before long. Even without the probability that the rate of ageing can eventually be slowed, so that it takes, say, 70 years to reach the equivalent of 60, society is beginning to move towards the acceptance of two succes-

CONNECTIONS

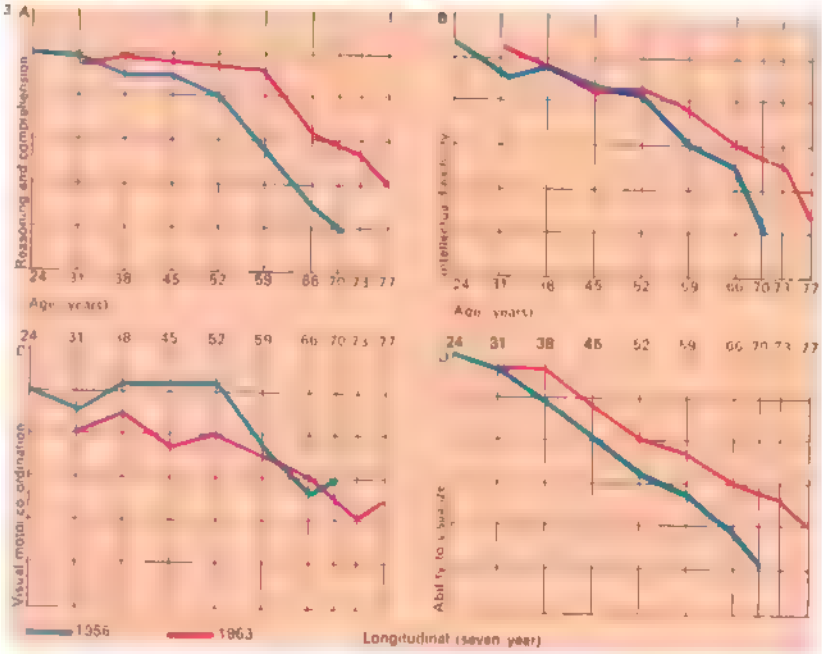
See also

1 Certain physical changes are characteristic of the ageing process. Hair greys or whitens as a result of hereditary factors. On the head, both men and women start to lose hair, although the process is more marked in men. Facial hair may increase, especially in women, but body hair thins or disappears. Height is reduced as spinal discs atrophy, leading to a sagging posture. Vision weakens at about 45 and a white or grey semicircle may later develop around the iris. Changes in the inner ear cause a loss of high-tone hearing. Smell and taste also become less sensitive. In men, a bulbous nose may develop due to faulty gland action. Dilation of blood vessels gives a spidery effect at the root of the nose. Teeth that are not looked after will be lost, leading to shrinkage of the jaw line. Under the chin and in the breasts, soft skin sags as a result of loss of the elastic protein collagen. The top layer of skin thins out all over the body so that blood vessels stand out, especially on the wrists. Skin also wrinkles and becomes discoloured. Senile freckles on the back of the hands and warts or raised red blood vessels on the body are common. Hand grip weakens and women may develop a swelling of the top finger joints. The heart is less able to respond to extra work as its valves grow more rigid and the body is more prone to illness because blood making capacity is reduced. Artery walls thicken. Lungs function less well due to reduced respiratory muscle power and distortion of the rib cage. The lungs' air spaces enlarge and breathlessness increases. Cellular units in the nervous system do not reproduce, so that nerve cells decline in number and the brain may lose bulk. The functions of the gastric juices, kidneys, spleen and pancreas deteriorate in effectiveness. Ovaries cease to function at the menopause. In men the testes atrophy, although sperm formation may continue to a late age. The prostate, a male gland, enlarges, tending to cause bladder trouble. Muscles lose bulk and become weaker. Bones grow lighter and more porous. Weight-bearing joints, particularly knees and ankles, may swell and become deformed. Osteoarthritis, a degenerative joint disease, is also common in spine and hips. Stiffness and loss of flexibility are increased by calcification in cartilage and ligaments. The shins may become tender and irritated as a result of the dryness of the skin that comes with old age. Corns and callouses grow thicker.





2 Behind the outward signs of ageing, several physiological processes [A, B] start to decline.



3 The popular belief that intelligence declines with age is challenged by recent research. When adults of various ages were tested at the same point in time the aged fared worse (red and blue lines). But generational differences may account for this because tests at different times on the same people revealed another picture (white lines). In reasoning and comprehension [A], intellectual flexibility [B] and ability to organize visual material [D] the same adults did better in 1963 than they had in 1956. Only their visual motor co-ordination [C] declined.

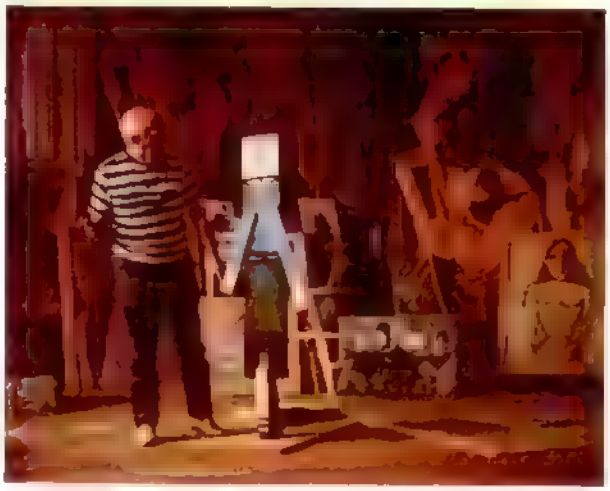
sive life-styles, with a renewal of education in middle and late life and a recognition of senior citizens" as a political force

Many people of both sexes already undergo an identity crisis very similar to a second adolescence at about the age of 45-50 in which they reassess their goals and achievements. This is because they can expect to live to the age of 70 to 80. Our ancestors dreamed of rejuvenation [5, 6,] but seldom lived long beyond 70. With the dissipation of myths about the inability of older people to learn or acquire new skills, senior education aimed at updating skills may be given higher priority. Destructive retirement practices which expect the retired to be idle, and legislation penalizing pensioners who earn, may come under increasing attack

Ageing does involve an increase in the incidence of illness, which may impair learning power and efficiency. Planning to meet the needs of old people must accommodate both the vigorous who want to remain fully engaged and a small but increasing proportion of those unable to do so, who need social and medical support. A similar range of options is needed in housing, some older people like to live in a mixed community, others with their contemporaries. The aim of social planning should be to make society "age blind" as well as colour blind so that status as a person is no more affected by age than by sex or race. While age inevitably lessens physical efficiency it does not as inevitably bring helplessness, dependence, foolishness or senility in its train

Psychological problems

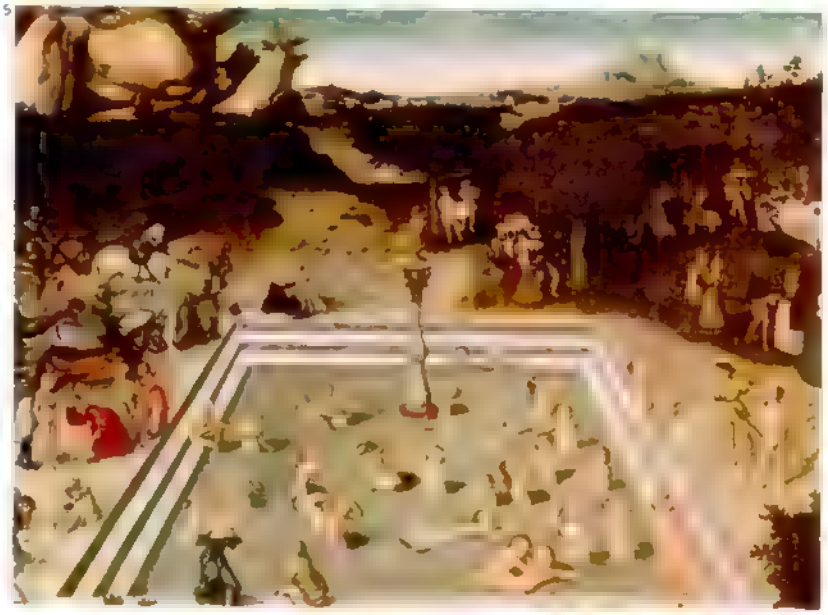
Of all the psychological problems that face the aged, the most difficult are perhaps isolation and the need for adaptability. In addition to physical handicaps and the diminishing number of social roles that are commonly available to the old, the aged have to cope with loss of the familiar – career, standard of living, surroundings and, above all, people. The gloomy search through the obituaries for yet another friend who has gone is bad enough, the death of a spouse is often a blow from which the aged do not recover. The ability to come to terms with these changes is vitally important



Creative power in old age has been demonstrated by many musicians, writers and artists, including Pablo Picasso (1881–1973). His longevity as a painter was surpassed only by that of Titian (c. 1487–1576), who produced some of his finest work in his 80s. Arturo Toscanini (1867–1957) was a vigorous conductor until the age of 87 while at the same age, Konrad Adenauer (1876–1967) was Chancellor of West Germany

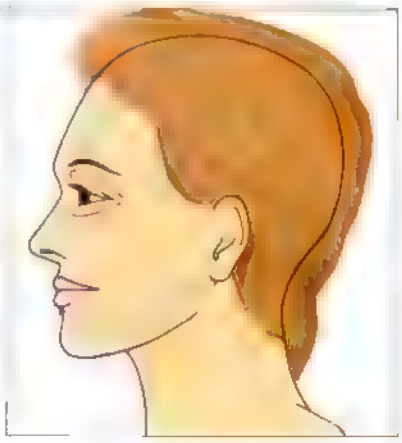
4 A Taoist believes that life can be prolonged by attaining the state of hsien. This can be reached by a combination of respiratory, dietary, sexual and gymnastic techniques. Man must inhale deeply when the earth breathes in, during the day. Grain and meat are impure so consumption is regulated and while sexual intercourse is not forbidden it involves a loss of vital body fluid

5 "The Fountain of Youth", by Lucas Cranach (1472–1553) is one of the many works that deal with the theme of rejuvenation. The idea that man could restore his youth by bathing in a magical river or spring occurs in many cultures. The Hindu Pool of Youth (700 BC) and the Hebrew River of Immortality are among the earliest examples. Alexander the Great supposedly searched for such a spring and Juan Ponce de Leon (1460–1521) was seeking it when he sighted the Florida peninsula unexpectedly in 1513. Rejuvenation was not thought impossible as snakes seemed to be reborn by shedding their skin and the mythical phoenix was reborn from fire



6 Satirizing hopes of rejuvenation, an eighteenth century artist envisaged a windmill that could grind up old women and reconstitute them as ladies of fashion. The most serious attempts to find an elixir of youth were made by alchemists who sought to gain power over nature through science. One of the basic themes of their search was that man could become eternal through association with things eternal. It was this theory that led Chinese alchemists to recommend that food should be eaten off golden plates. For in this way the patient would take in a little of the metal each time he ate and since gold was incorruptible he would gradually achieve a higher state of being

7 Plastic surgery is a drastic means of "rejuvenating" the face that has grown in favour in recent years. Sagging and wrinkled skin, brought about by a decline in the elasticity of muscles is stretched back from the face and neck by a surgeon who then removes the excess and stitches the gap beneath the hairline. The process leaves the skin tauter but often has to be repeated



Death, grieving and loss

From the moment of birth each of us begins to develop a model of the world in our imagination. This model includes everything that we know, or think we know, of the world around us, it includes our own bodies and minds insofar as we can view these ourselves, it includes the things and the people we know, and our plans, expectations and hopes [1]. Sometimes, events will occur that invalidate a part or parts of our world model – the unexpected loss of a loved one, for example – and we may be faced with the need to abandon many of our assumptions and painfully to rebuild the world it has taken so long to create. The experience of a change of this kind has been termed a "psychosocial transition" or crisis period.

Reactions to change and loss

Periods of crisis are both time and energy consuming and they follow a more or less consistent sequence [2]. At first, particularly if the change is sudden or massive, there is likely to be a stage of shock, denial or disbelief. The individual is unable to take in the reality of what is happening and tends to behave as if no change were occurring at all.

Before long, however, realization begins to dawn and there is a period of striving in the course of which the person bitterly tries to recover the world that he or she is losing and to preserve the old model. Repeated frustration causes him or her to abandon the struggle and a period of disillusionment, apathy and despair sets in. Finally, little by little, the appetites of life return and new beginnings are made so that the final phase is one of reorganization and recovery.

Death and bereavement

The most devastating and overwhelming type of change is possibly that which occurs when a previously healthy individual develops a fatal illness. Death implies so radical a change that not only is the dying person likely to avoid facing the prospect of his or her own death but everyone around tends to collude in the pretence that the illness will disappear. Remarkably, given emotional support and the relief of pain and other symptoms, many people do eventually express a wish to be told the full facts of their illness. They will often arrive at a stage of acceptance in which they fully realize the implications of the illness and can reorganize and enjoy to the full the life that remains [6, 7].

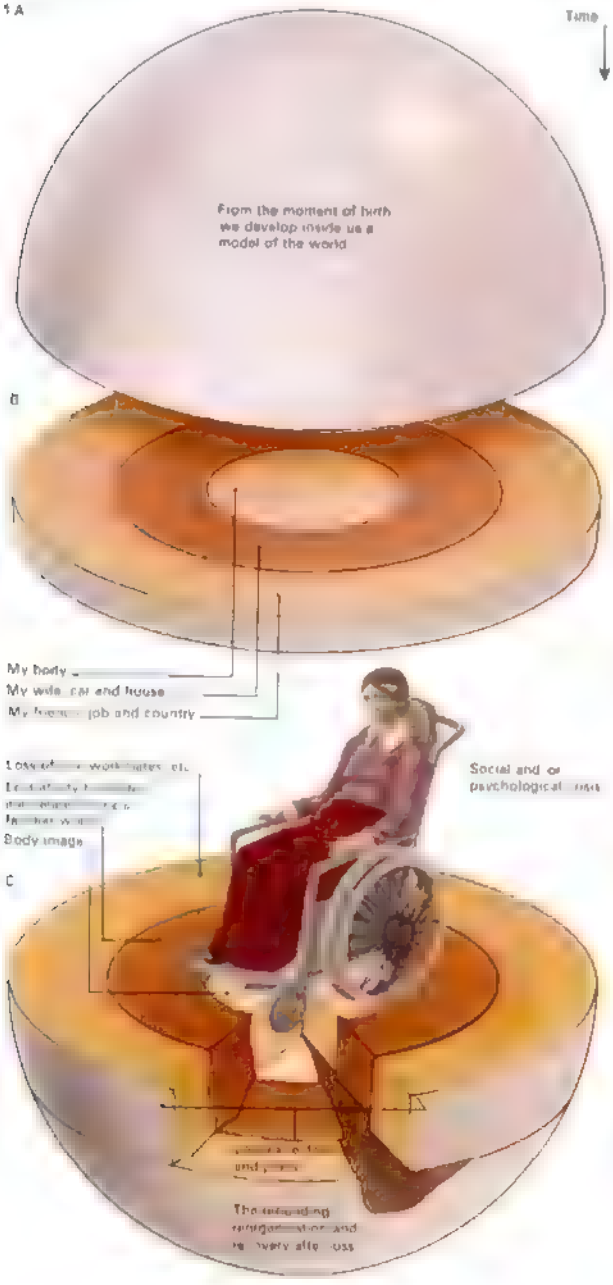
Coping with death implies the possession of a philosophy of life that includes death as a valid part of the world model. In modern Western society this is rare and death has taken the place of sex as a taboo topic. Consequently the dying and the bereaved are sometimes deprived of human relationships at the time when they need them most.

It would be fair to assume that once a person has died his or her troubles are over. For those who remain, however, a fresh stage of adjustment may be just beginning. They may be helped or hindered in this adjustment by the rituals and social events that, in all societies, mark the transition from life to death. The survivor is faced with the need to undertake a major change in his or her model of the world. It is hard to deny the fact of death for long and soon the phase of shock is replaced by the pangs of grief – characterized by intense pining.

Some people report an illusory sense of

CONNECTIONS

See also



1 The model of the world that we contain within us can be pictured as an onion whose centre point is our view of our selves. Around this core are our bodies and those parts of the world that are most intimately ours, and beyond them aspects that are more widely shared: the horizontal slice (B) representing "now" and the upper part of the figure "the remembered past" (A). Any event, such as a disabling accident (C), which renders a large part of this world model incorrect or useless gives rise to a difficult period of crisis. As a result of this transition a major and painful restructuring of many layers of the onion is necessary.

2 Death, the most fundamental period of transition in our lives, is a crisis that many people in Western society face unprepared. This lack of preparation is reflected in the way that, all too often, relatives, doctors and nurses try to hide the fact of death and maintain the pretence that a dying person is not very ill (A). When a person dies those closest to him pass through the grief (B) which for a while will disorganize their lives and make it hard for them to support each other. If they face up to their grief and receive encouragement and understanding from others they will be able to cope better with the successive losses and griefs of old age (C).

the continuing presence of the dead person that may be retained for many years, but in time the facts are accepted and, if all goes well, the bereaved person discovers a fresh identity. In other types of loss, such as that following divorce, there may be a similar sequence of reactions but these are more likely to be complicated by anger and self-reproach.

Disablement and old age

Illness and accident may give rise to physical disablement. The patient is then faced with the need to grieve not only for the lost ability but also for those aspects of life that went with it. Thus the person who loses a limb grieves for the job that he or she can no longer carry out, for the lost sports and other activities and for the future prospects that seem to have disappeared.

Disabled people need time and emotional support if they are to come through the grief that must be expressed for the real losses they have undergone. They may be reluctant to learn the skills appropriate to their disability until this process of realization is complete.

Old people tend to suffer a succession of bereavements and disablements. They often attempt to cope with these by narrowing their horizons and disengaging themselves from many of the people around them. But at the same time they become vulnerable to changes in the small environment that is all they have left. It is important to recognize that an old person's room and its familiar clutter of objects are much more important than they would be to a younger person.

Periods of grief and loss are times of danger but they are also times of opportunity. Those who come through them may emerge more mature and secure than before. Those who succumb – by excessive avoidance or by "caving in" – may find future losses even more hard to take. The loving support and care of close friends and relatives remains the most valuable consolation available to those facing death or loss. Modern society suffers from an increasing embarrassment about the expression of feeling. Spontaneous gestures of affection are the gifts sometimes hardest to give but most valued of all because they tell the bereaved that he or she is not alone.

KEY

The expression of grief in the human face reflects a deep-seated need to cry. This has a "signal" function and evokes support in others. But knowing that it is useless to search for the lost person, efforts are often made to inhibit grief, with varying success and inner tension.

3

3 The rituals attending death have usually been seen as support to the dead, but most of them also have a psychological function for the living. The provision of food, drink and familiar possessions, as in this Bronze Age Danish burial, comfort the survivors as well.

4

4 In the former Indian tradition of "suttee" the widow of the dead man joined her husband on the funeral pyre. In this culture she was treated as one of the familiar possessions that accompanied him into the next world. It was also assumed, perhaps, that life for the survivor would, or certainly should, be intolerable.

5

5 Societies vary greatly in the ritual expression of mourning. In this Chinese funeral the mourners wear white and the colourful coffin seems to reflect the hope of a contented future for the deceased. The sarcophagus bears the inscription "good luck" in Chinese characters. But this does not prevent grief from being expressed and professional wailers may be employed by the family to give social sanction to the shedding of tears among relatives and friends.

6

6 In facing the prospect of his or her own death a person can come through the grief and become free to enjoy the life that remains to him or her. This woman, knowing full well that she was within a few days of death, was yet able to enjoy a ride in the garden of a hospital.

7 A

B

7 These pictures were drawn by a woman with terminal cancer. Her feelings when she first entered hospital [A] contrast sharply with her outlook a few weeks later [B]. The pictures reflect a remarkable transformation from all-consuming pain to a balanced, peaceful frame of mind, even though staff made no attempt to conceal from this patient the facts of her fatal illness.

Questions of life and death

Man has always asked questions about what he ought to do, where his duty lies and what rights he possesses or grants to others. He is a moral being because he cannot evade choices. Since these choices are often difficult to make, man has evolved systems of principles or morals – usually as expressed by a religion or philosophy – in order to help him make his choices and measure his actions. In this way he has some concrete and shared idea of what good and evil mean.

The question remains, particularly in a largely secular society, as to whether these principles are absolute – given, as it were, by some external and superior force – or whether they are man-made. Fyodor Dostoevsky (1821–81) asserted that "If God does not exist, everything is permitted", but it may be that our principles represent a consensus of rules that make social life possible.

Attitudes to moral systems can be classified broadly into two types. Moral principles (for example, "Thou shalt not kill") can be formulated and then applied whatever the consequences and regardless of the specific individual case. (Few if any values, however, have been so applied – men have been encouraged by many churches to kill in war, for instance.) In contrast to this "absolutist" approach to morality, a "situationist" or "utilitarian" approach to moral values is more usual today, whereby the particular circumstances of each case are considered.

Science and morality

There is nothing really new about any present-day moral problems, but they are now perhaps more pressing than in previous ages because modern science has greatly increased the scope and range of the consequences of man's decisions.

This is most obvious in modern warfare. Combatants are now far removed from each other and, in the process, distanced from the consequences of their actions. War is also no longer simply a combat between armies: the humane distinction between combatants and non-combatants has been obliterated. This raises the question of whether there can be any "innocent bystanders" in modern conflicts. War has become total, limited only by the fear of universal destruction.

The problem is at its most extreme with the atomic and hydrogen bombs [1]. Many people have questioned the development and production of these bombs, let alone their use, both because their destructiveness outweighs any justification for their use and because of the probable long-term effects of radiation. Does the end of "saving" lives justify the horrific means of attaining that saving? Even the peaceful use of nuclear power carries the risk of lethal pollution of the atmosphere for many generations. Behind the practical problems of using it peacefully is the moral one of whether it is possible to balance the benefits, however great, against the risk, however small and remote it may in fact be.

Dilemmas in medicine

Advances in medical science pose problems in unusual guises, from the morality of using animals in experiments [4] to the question of exactly when "spare parts" for transplant surgery [3] may be removed from "dead" donors in the interests of another human life.

Another area of medical progress –

CONNECTIONS

See also
Philosophy and ethics
Moral development
Structure of societies
Rule of law
Preventive medicine
Birth control
Development of surgery
Behaviour therapy

1

1 Atomic weapons raise serious moral problems for man. In 1945, the first two atomic bombs totally destroyed Hiroshima and Nagasaki in Japan. The US president, Harry S. Truman, explained that if the bombs had not been dropped, the war would have dragged on for 18 months longer at a probable cost of two million Japanese and one million American lives. Ending war swiftly and saving lives are unarguably moral intentions for a head of state, but this situation may have been a prime example of the way moral dilemmas defy reduction to practical terms. The "equation of suffering" on which Truman based his argument does not necessarily provide a moral justification for his action.

2

3

2 The use of gas in World War I was abandoned not for moral reasons, but because it proved to be an unreliable weapon. Other chemical methods have since been used, although germ warfare has not. In many instances, although by no means all, there seem, fortunately, to be "natural sanctions" that set limits to human behaviour.

3 The first heart transplant was made by Christiaan Barnard (1922–) in 1966. Although a great medical feat, the problem remains whether large sums and rare skills should be used to save a few lives while less spectacular operations are neglected. However, without such pioneering operations medical knowledge cannot advance.

increased control over reproduction – has given new dimensions to one of the most intractable of moral questions: who has the right to decide in matters of life and death? The broad question of social responsibility and individual rights is also relevant to another area of birth control – "genetic engineering" and eugenics. For the scientist could eventually be faced with the morality of "improving the breed" at the expense of a couple's right to produce children.

The morality of preventing life – by sterilization, prohibition of reproduction on medical grounds, or abortion [5] – is closely linked with that of taking life. Voluntary euthanasia (mercy killing), for example, in the face of intolerable and incurable pain or because life has been reduced to vegetable existence, has been claimed as a fundamental human right. In such cases the conflicting moral arguments of the "absolutists" and "relativists" are thrown into sharpest relief.

Sometimes moral problems arise because the state wants to impose its will on the individual. Even if it stops short of putting down undesirables [Key], has society the right to order the sterilization of a mentally retarded girl whose children are likely to be handicapped? Should it reform habitual criminals by aversion therapy or, more radically, by a brain operation that makes them irrevocably docile even if the result is a pacified robot rather than a human being? The possibility of such methods being used by an unscrupulous or frightened state in order to control political dissidents is a daunting one.

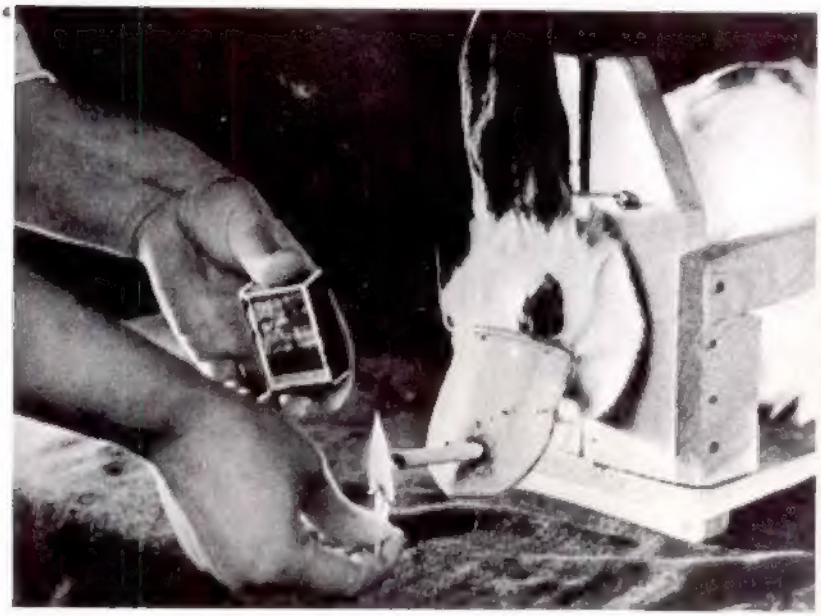
The roots of morality

All moral traditions, both religious and secular, find such radical measures repugnant because these traditions make man the norm of morality and hold that one may not abuse the dignity of another person without harming oneself. The whole of humanity is, ultimately, interdependent and interrelated, and men are responsible for each other and for future generations. In borderline situations there will always be conflicts of duties, but love and respect for other human beings, truthfulness and honesty in relationships provide, if not final answers, a starting-point on which to base any moral system.

KEY

Capital punishment, here summarily administered in Manchu China, raises the fundamental question of whether life should be taken for any reason at all. Is it an effective deterrent? What sort of crimes should it be imposed for: treason, premeditated murder, acts of terrorism? Because the moral and practical arguments are so difficult and provoke such emotion most societies invest the power over life and death not in the individual but in "justice", "God" or "the people".

4



4 Animals are still widely used in medical research. The justification is that the lives and health of humans are more important than those of animals: that it is better to find out the ill effects of drugs or smoking on rabbits than on humans. Does this justify the torture of animals? The balance between benefit to man and intolerable harm to animals is difficult to establish morally.

5 Human characteristics are obvious in a 12-week-old fetus. Its genetic coding has already dictated the colour of its eyes and the pattern of its fingerprints, but it is still totally dependent on its mother for oxygen and nourishment. These conflicting factors invite opposing arguments: either that the fetus is simply a part of the mother's body; or that the fetus is to be considered as a human life. Abortion thus becomes, by the first argument, justified if used to preserve the health of the mother or prevent the suffering of the child; and by the second, an indefensible act of murder.

5

6

6 Law and order are necessary in all societies, but force should, in theory, be used only to prevent a major breach of the peace. The aim of these not policemen, once they have to be called in, should be to disperse the crowd without causing injury. The right of assembly and peaceful demonstration is basic to all free societies and the erosion of that right one of the first signs of repression.

Man at war –
A South Vietnamese soldier holds his son's body, tagged for burial.